Lecture Notes in Electrical Engineering 90

Sio Iong Ao
Len Gelman
Editors

Electrical Engineering and Applied Computing

Springer

VISIT…

Lecture Notes in Electrical Engineering

Volume 90

For further volumes:
http://www.springer.com/series/7818

Sio Iong Ao · Len Gelman
Editors

Electrical Engineering and Applied Computing

Springer

Editors
Sio Iong Ao
International Association of Engineers
Unit 1, 1/F, 37-39 Hung To Road
Kwun Tong
Hong Kong
e-mail: siao@graduate.hku.hk

Len Gelman
Applied Mathematics and Computing
School of Engineering
Cranfield University
Cranfield
UK
e-mail: l.gelman@cranfield.ac.uk

ISSN 1876-1100 e-ISSN 1876-1119

ISBN 978-94-007-1191-4 e-ISBN 978-94-007-1192-1

DOI 10.1007/978-94-007-1192-1

Springer Dordrecht Heidelberg London New York

© Springer Science+Business Media B.V. 2011

No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Cover design: eStudio Calamar, Berlin/Figueres

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

A large international conference in Electrical Engineering and Applied Computing was held in London, U.K., 30 June–2 July, 2010, under the World Congress on Engineering (WCE 2010). The WCE 2010 was organized by the International Association of Engineers (IAENG); the Congress details are available at: http://www.iaeng.org/WCE2010. IAENG is a non-profit international association for engineers and computer scientists, which was founded originally in 1968. The World Congress on Engineering serves as good platforms for the engineering community to meet with each other and exchange ideas. The conferences have also struck a balance between theoretical and application development. The conference committees have been formed with over two hundred members who are mainly research center heads, faculty deans, department heads, professors, and research scientists from over 30 countries. The conferences are truly international meetings with a high level of participation from many countries. The response to the Congress has been excellent. There have been more than one thousand manuscript submissions for the WCE 2010. All submitted papers have gone through the peer review process, and the overall acceptance rate is 57%.

This volume contains fifty-five revised and extended research articles written by prominent researchers participating in the conference. Topics covered include Control Engineering, Network Management, Wireless Networks, Biotechnology, Signal Processing, Computational Intelligence, Computational Statistics, Internet Computing, High Performance Computing, and industrial applications. The book offers the state of the art of tremendous advances in electrical engineering and applied computing and also serves as an excellent reference work for researchers and graduate students working on electrical engineering and applied computing.

Sio Iong Ao
Len Gelman

Contents

1 **Mathematical Modelling for Coal Fired Supercritical Power Plants and Model Parameter Identification Using Genetic Algorithms** 1
Omar Mohamed, Jihong Wang, Shen Guo, Jianlin Wei, Bushra Al-Duri, Junfu Lv and Qirui Gao

2 **Sequential State Computation Using Discrete Modeling** 15
Dumitru Topan and Lucian Mandache

3 **Detection and Location of Acoustic and Electric Signals from Partial Discharges with an Adaptative Wavelet-Filter Denoising** 25
Jesus Rubio-Serrano, Julio E. Posada and Jose A. Garcia-Souto

4 **Study on a Wind Turbine in Hybrid Connection with a Energy Storage System** 39
Hao Sun, Jihong Wang, Shen Guo and Xing Luo

5 **SAR Values in a Homogenous Human Head Model** 53
Levent Seyfi and Ercan Yaldız

6 **Mitigation of Magnetic Field Under Overhead Transmission Line** 67
Adel Zein El Dein Mohammed Moussa

7 **Universal Approach of the Modified Nodal Analysis for Nonlinear Lumped Circuits in Transient Behavior** 83
Lucian Mandache, Dumitru Topan and Ioana-Gabriela Sirbu

8 Modified 1.28 Tbit/s (32 × 4 × 10 Gbit/s) Absolute Polar Duty Cycle Division Multiplexing-WDM Transmission Over 320 km Standard Single Mode Fiber 95
Amin Malekmohammadi

9 Wi-Fi Wep Point-to-Point Links 105
J. A. R. Pacheco de Carvalho, H. Veiga, N. Marques, C. F. Ribeiro Pacheco and A. D. Reis

10 Interaction Between the Mobile Phone and Human Head of Various Sizes 115
Adel Zein El Dein Mohammed Moussa and Aladdein Amro

11 A Medium Range Gbps FSO Link 125
J. A. R. Pacheco de Carvalho, N. Marques, H. Veiga, C. F. Ribeiro Pacheco and A. D. Reis

12 A Multi-Classifier Approach for WiFi-Based Positioning System 135
Jikang Shin, Suk Hoon Jung, Giwan Yoon and Dongsoo Han

13 Intensity Constrained Flat Kernel Image Filtering, a Scheme for Dual Domain Local Processing 149
Alexander A. Gutenev

14 Convolutive Blind Separation of Speech Mixtures Using Auditory-Based Subband Model 161
Sid-Ahmed Selouani, Yasmina Benabderrahmane, Abderraouf Ben Salem, Habib Hamam and Douglas O'Shaughnessy

15 Time Domain Features of Heart Sounds for Determining Mechanical Valve Thrombosis 173
Sabri Altunkaya, Sadık Kara, Niyazi Görmüş and Saadetdin Herdem

16 On the Implementation of Dependable Real-Time Systems with Non-Preemptive EDF 183
Michael Short

17 Towards Linking Islands of Information Within Construction Projects Utilizing RF Technologies 197
Javad Majrouhi Sardroud and Mukesh Limbachiy

18 A Case Study Analysis of an E-Business Security Negotiations Support Tool 209
Jason R. C. Nurse and Jane E. Sinclair

19 Smart Card Web Server 221
Lazaros Kyrillidis, Keith Mayes and Konstantinos Markantonakis

20 A Scalable Hardware Environment for Embedded Systems Education 233
Tiago Gonçalves, A. Espírito-Santo, B. J. F. Ribeiro and P. D. Gaspar

21 Yield Enhancement with a Novel Method in Design of Application-Specific Networks on Chips 247
Atena Roshan Fekr, Majid Janidarmian, Vahhab Samadi Bokharaei and Ahmad Khademzadeh

22 On-Line Image Search Application Using Fast and Robust Color Indexing and Multi-Thread Processing 259
Wichian Premchaisawadi and Anucha Tungkatsathan

23 Topological Mapping Using Vision and a Sparse Distributed Memory 273
Mateus Mendes, A. Paulo Coimbra and Manuel M. Crisóstomo

24 A Novel Approach for Combining Genetic and Simulated Annealing Algorithms 285
Younis R. Elhaddad and Omar Sallabi

25 Buyer Coalition Formation with Bundle of Items by Ant Colony Optimization 297
Anon Sukstrienwong

26 Coevolutionary Grammatical Evolution for Building Trading Algorithms 311
Kamal Adamu and Steve Phelps

27 High Performance Computing Applied to the False Nearest Neighbors Method: Box-Assisted and kd-Tree Approaches 323
Julio J. Águila, Ismael Marín, Enrique Arias, María del Mar Artigao and Juan J. Miralles

28 Ethernet Based Implementation of a Periodic Real Time Distributed System 337
Sahraoui Zakaria, Labed Abdennour and Serir Aomar

29 Preliminary Analysis of Flexible Pavement Performance Data Using Linear Mixed Effects Models 351
Hsiang-Wei Ker and Ying-Haur Lee

30 Chi-Squared, Yule's Q and Likelihood Ratios in Tabular Audiology Data 365
Muhammad Naveed Anwar, Michael P. Oakes and Ken McGarry

31 Optimising Order Splitting and Execution with Fuzzy Logic Momentum Analysis 377
Abdalla Kablan and Wing Lon Ng

32 The Determination of a Dynamic Cut-Off Grade for the Mining Industry 391
P. V. Johnson, G. W. Evatt, P. W. Duck and S. D. Howell

33 Improved Prediction of Financial Market Cycles with Artificial Neural Network and Markov Regime Switching 405
David Liu and Lei Zhang

34 Fund of Hedge Funds Portfolio Optimisation Using a Global Optimisation Algorithm 419
Bernard Minsky, M. Obradovic, Q. Tang and Rishi Thapar

35 Increasing the Sensitivity of Variability EWMA Control Charts 431
Saddam Akber Abbasi and Arden Miller

36 Assessing Response's Bias, Quality of Predictions, and Robustness in Multiresponse Problems 445
Nuno Costa, Zulema Lopes Pereira and Martín Tanco

37 Inspection Policies in Service of Fatigued Aircraft Structures 459
Nicholas A. Nechval, Konstantin N. Nechval and Maris Purgailis

38 Toxicokinetic Analysis of Asymptomatic Hazard Profile of Welding Fumes and Gases 473
Joseph I. Achebo and Oviemuno Oghoore

39 Classification and Measurement of Efficiency and Congestion of Supply Chains 487
Mithun J. Sharma and Song Jin Yu

40 Comparison of Dry and Flood Turning in Terms of Dimensional Accuracy and Surface Finish of Turned Parts 501
Noor Hakim Rafai and Mohammad Nazrul Islam

41 Coordinated Control Methods of Waste Water Treatment Process 515
Magdi S. Mahmoud

42 Identical Parallel-Machine Scheduling and Worker Assignment Problem Using Genetic Algorithms to Minimize Makespan 529
Imran Ali Chaudhry and Sultan Mahmood

43 Dimensional Accuracy Achievable in Wire-Cut Electrical Discharge Machining 543
Mohammad Nazrul Islam, Noor Hakim Rafai and Sarmilan Santhosam Subramanian

44 Nash Game-Theoretic Model for Optimizing Pricing and Inventory Policies in a Three-Level Supply Chain 555
Yun Huang and George Q. Huang

45 Operating Schedule: Take into Account Unexpected Events in Case of a Disaster 567
Issam Nouaouri, Jean Christophe Nicolas and Daniel Jolly

46 Dynamic Hoist Scheduling Problem on Real-Life Electroplating Production Line 581
Krzysztof Kujawski and Jerzy Świątek

47 Effect of HAART on CTL Mediated Immune Cells: An Optimal Control Theoretic Approach 595
Priti Kumar Roy and Amar Nath Chatterjee

48 Design, Development and Validation of a Novel Mechanical Occlusion Device for Transcervical Sterilization 609
Muhammad Rehan, James Eugene Coleman and Abdul Ghani Olabi

49 Investigation of Cell Adhesion, Contraction and Physical Restructuring on Shear Sensitive Liquid Crytals 623
Chin Fhong Soon, Mansour Youseffi, Nick Blagden and Morgan Denyer

50 On the Current Densities for the Electrical Impedance Equation 637
Marco Pedro Ramirez Tachiquin, Jose de Jesus Gutierrez Cortes, Victor Daniel Sanchez Nava and Edgar Bernal Flores

51 Modelling of Diseased Tissue Diffuse Reflectance and Extraction of Optical Properties 649
Shanthi Prince and S. Malarvizhi

52 Vertical Incidence Increases Virulence in Pathogens: A Model Based Study 661
Priti Kumar Roy, Jayanta Mondal and Samrat Chatterjee

53 Chaotic Oscillations in Hodgkin–Huxley Neural Dynamics 675
Mayur Sarangdhar and Chandrasekhar Kambhampati

54 Quantification of Similarity Using Amplitudes and Firing Times of a Hodgkin–Huxley Neural Response 687
Mayur Sarangdhar and Chandrasekhar Kambhampati

55 Reduction of HIV Infection that Includes a Delay with Cure Rate During Long Term Treatment: A Mathematical Study 699
Priti Kumar Roy and Amar Nath Chatterjee

Chapter 1
Mathematical Modelling for Coal Fired Supercritical Power Plants and Model Parameter Identification Using Genetic Algorithms

Omar Mohamed, Jihong Wang, Shen Guo, Jianlin Wei, Bushra Al-Duri, Junfu Lv and Qirui Gao

Abstract The paper presents the progress of our study of the whole process mathematical model for a supercritical coal-fired power plant. The modelling procedure is rooted from thermodynamic and engineering principles with reference to the previously published literatures. Model unknown parameters are identified using Genetic Algorithms (GAs) with 600MW supercritical power plant on-site measurement data. The identified parameters are verified with different sets of measured plant data. Although some assumptions are made in the modelling process to simplify the model structure at a certain level, the supercritical

O. Mohamed (✉) · J. Wang · S. Guo · J. Wei
School of Electrical, Electronics, and Computer Engineering, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
e-mail: ORM@bham.ac.uk

J. Wang
e-mail: j.h.wang@bham.ac.uk

S. Guo
e-mail: s.guo@bham.ac.uk

J. Wei
e-mail: j.wei@bham.ac.uk

B. Al-Duri
School of Chemical Engineering, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
e-mail: B.Al-Duri@bham.ac.uk

J. Lv · Q. Gao
Department of Thermal Engineering, Tsinghua University, Beijing, People's Republic of China
e-mail: lvjf@tsinghua.edu.cn

Q. Gao
e-mail: gaoqr@tsinghua.edu.cn

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_1,
© Springer Science+Business Media B.V. 2011

coal-fired power plant model reported in the paper can represent the main features of the real plant once-through unit operation and the simulation results show that the main variation trends of the process have good agreement with the measured dynamic responses from the power plants.

Nomenclature

ff	Fitness function for genetic algorithms
ffr	Pulverized fuel flow rate (kg/s)
h	Enthalpy per unit mass (MJ/kg)
K	Constant parameter
k	Mass flow rate gain
m	Mass (kg)
$\dot{m}$	Mass flow rate (kg/s)
P	Pressure of a heat exchanger (MPa)
$\dot{Q}$	Heat transfer rate (MJ/s)
R	Response
T	Temperature (°C)
t	Time (s)
τ	Time constant (s)
U	Internal energy (MJ)
V	Volume of fluid (m^3)
$\dot{W}$	Work rate or power (MW)
x	Generator reactance (p.u)
y	Output vector
ρ	Density (kg/m^3)
χ	Valve opening
δ	Rotor angle (rad)
θ	Mechanical angle (rad)
ω	Speed (p.u)
Γ	Torque (p.u)

Subscripts

a	Accelerating
air	Air
e	Electrical
d	Direct axis
ec	Economizer
hp	High pressure turbine
hx	Heat exchanger
i	Inlet
ip	Intermediate pressure turbine
me	Mechanical
ms	Main steam
m	Measured

o	Outlet
out	Output of the turbine
q	Quadrature axis
rh	Reheater
sh	Superheater
si	Simulated
ww	Waterwall

Abbreviations

BMCR	Boiler maximum continuous rate
ECON	Economizer
GA	Genetic algorithm
HP	High pressure
HX	Heat exchanger
IP	Intermediate pressure
MS	Main steam
RH	Reheater
SC	Supercritical
SH	Superheater
WW	Waterwall

1.1 Introduction

The world is now facing the challenge of the issues from global warming and environment protection. On the other hand, the demand of electricity is growing rapidly due to economic growth and increases in population, especially in the developing countries, for example, China and India. With the consideration of environment and sustainable development in energy, renewable energy such as wind, solar, and tidal wave should be only resources to be explored in theory. But the growth in demand is also a heavy factor in energy equations so the renewable energy alone is unlikely able to generate sufficient electricity to fill the gap in the near future. Power generation using fossil fuels is inevitable, especially, coal fired power generation is found to be an unavoidable choice due to its huge capacity and flexibility in load following. As a well know fact, the conventional coal fired power plants have a huge environmental impact and lower energy conversion efficiencies. Any new coal fired power plants must be cleaner with more advanced and improved technologies.

Apart from Carbon Capture and Storage, supercritical power plants might be the most suitable choice with consideration of the factors in environmental enhancement, higher energy efficiency and economic growth. However, there has

been an issue to be addressed in its dynamic responses and performance in relation with conventional subcritical plants due to the difference in the process structure and energy storage drum [1]. The characteristics of supercritical plants require the considerable attention and investigation. Supercritical boilers have to be once-through type boilers because there is not distinction between water and steam phases in supercritical process so there is no drum to separate water steam mixture. Due to the absence of the drum, the once-through boilers have less stored energy and faster responses than the drum-type boilers. There are several advantages of supercritical power plants [2, 3] over traditional subcritical plants include:

- Higher cycle efficiency (Up to 46%) and lower fuel consumption.
- Reduced CO_2 emissions per unit power generation.
- Be fully integratable with CO_2 capture technology.
- Fast load demand following (in relatively small load demand changes).

However, some concerns are also raised in terms of its dynamic responses with regards to the demand for dynamic response speed. This is mainly caused by its once-through structure, that is, there is no drum to store energy as a buffer to response rapid changes in load demand.

The paper is to develop a mathematical model for the whole plant process to study dynamic responses aiming at answering the questions in dynamic response speed. From the literature survey, several models have been reported with emphasis on different aspects of the boiler characteristics. Studying the dynamic response and control system of once-through supercritical (SC) units can be traced back to 1958 when work was done on a time-based simulation for Eddystone I unit of Philadelphia Electric Company and the work was extended for simulation of Bull run SC generation unit later in 1966 [4].

Yutaka Suzuki et al. modelled a once through SC boiler in order to improve the control system of an existing supercritical oil-fired plant. The model was based on nonlinear partial differential equations, and the model was validated through simulation studies [5]. Wataro Shinohara et al. presented a simplified state space model for SC once through boiler-turbine system and designed a nonlinear controller [6]. Pressure node model description was introduced by Toshio Inoue et al. for power system frequency simulation studies [7]. Intelligent techniques contributions have yielded an excellent performance for modeling. Neural network has been introduced to model the SC power plant with sufficiently accurate results if they are trained with suitable data provided by operating unit [8]. However, neural network performances are unsatisfactory to simulate some emergency conditions of the plant because NN method depends entirely on the data used for the learning process, not on physical laws. Simulation of SC boilers may be achieved either theoretically based on physical laws or empirically based on experimental work. In this paper, the proposed mathematical model is based on thermodynamic principles and the model parameters are identified by using the data obtained from a 600MW SC power plant [9]. The simulation results show that the model is trustable to simulate the whole once-through mode of operation at a certain level of accuracy.

1.2 Mathematical Model of the Plant

1.2.1 Plant Description

The unit of a once-through supercritical 600MW power plant is selected for the modelling study. The schematic view of the boiler is shown in Fig. 1.1. Water from the feedwater heater is heated in the economizer before entering the superheating stages through the waterwall. The superheater consists of three sections which are low temperature superheater, platen superheater, and final stage superheater. The main outlet steam temperature is about 571°C at the steady state and a pressure is 25.5 MPa. There are 2 reheating sections in the boiler for reheating the steam exhausted from the high pressure turbine. The inlet temperature of the reheater is 309°C and the outlet temperature is nearly 571°C and average pressure is 4.16 MPa. The reheated steam is used to energize the intermediate pressure turbine. The mechanical power is generated through multi-stage turbines to provide an adequate expansion of the steam through the turbine and subsequently high thermal efficiency of the plant.

1.2.2 Assumptions Made for Modelling

Assumptions are made to simplify the process which should be acceptable by plant engineers and sufficient to transfer the model from its complex physical model to lead to simple mathematical model for the research purpose. Some of these assumptions are usually adopted for modelling supercritical or subcritical boilers [10]. Modelling in the work reported in the paper, the following general assumptions are made:

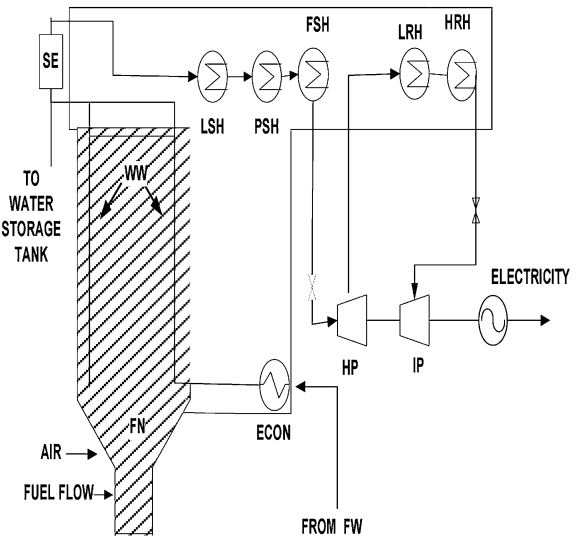


Fig. 1.1 Schematic view of the plant

- Fluid properties are uniform at any cross section, and the fluid flow in the boiler tubes is one-phase flow.
- In the heat exchanger, the pipes for each heat exchanger are lumped together to form one pipe.
- Only one control volume is considered in the waterwall.
- The dynamic behaviour of the air and gas pressure is neglected.

1.2.3 The Boiler Model

1.2.3.1 Heat Exchanger Model

The various heat exchangers in the boiler are modelled by the principles of mass and energy balances. The sub-cooled water in the economizer is transferred directly to a supercritical steam through the waterwall without passing the evaporation status. The equations are converted in terms of the derivatives (or variation rates) pressure and temperature of the heat exchanger. The mass balance equation of the heat exchanger (control volume) is:

$$\frac{dm}{dt} = \dot{m}_i - \dot{m}_o \tag{1.1}$$

For the constant effective volume, Eq. 1.1 will be:

$$V\frac{d\rho}{dt} = \dot{m}_i - \dot{m}_o$$

The density is a differentiable function of two variables which can be the temperature and pressure inside the control volume, thus we have:

$$V\left(\left.\frac{\partial\rho}{\partial P}\right|_T \cdot \frac{dP}{dt} + \left.\frac{\partial\rho}{\partial T}\right|_P \cdot \frac{dT}{dt}\right) = \dot{m}_i - \dot{m}_o$$

The energy balance equation:

$$\frac{dU_{hx}}{dt} = \dot{Q}_{hx} + \dot{m}_i h_i - \dot{m}_o h_o$$

Also,

$$\begin{aligned}\frac{dU_{hx}}{dt} = {} & V\left[h\left(\left.\frac{\partial\rho}{\partial P}\right|_T \cdot \frac{dP}{dt} + \left.\frac{\partial\rho}{\partial T}\right|_P \cdot \frac{dT}{dt}\right) + \rho\left(\left.\frac{\partial \mathrm{h}}{\partial P}\right|_T \cdot \frac{dP}{dt} + \left.\frac{\partial h}{\partial T}\right|_P \cdot \frac{dT}{dt}\right)\right] \\ & - V\frac{dP}{dt}V\left[h\left(\left.\frac{\partial\rho}{\partial P}\right|_T \cdot \frac{dP}{dt} + \left.\frac{\partial\rho}{\partial T}\right|_P \cdot \frac{dT}{dt}\right) + \rho\left(\left.\frac{\partial h}{\partial P}\right|_T \cdot \frac{dP}{dt} + \left.\frac{\partial h}{\partial T}\right|_P \cdot \frac{dT}{dt}\right)\right] \\ & - V\frac{dP}{dt}\dot{Q}_{hx} + \dot{m}_i h_i - \dot{m}_o h_o \end{aligned} \tag{1.2}$$

Combining (1.1) and (1.2) to get the pressure and temperature state derivatives,

$$\dot{P} = \frac{\dot{Q}_{hx} + \dot{m}_i H_i - \dot{m}_o H_o}{\tau} \tag{1.3}$$

$$\dot{T} = C(\dot{m}_i - \dot{m}_o) - D\dot{P} \tag{1.4}$$

Where:

$$H_i = \left(h_i - h - \frac{\rho \frac{\partial h}{\partial T}\Big|_P}{\frac{\partial \rho}{\partial T}\Big|_P} \right) \tag{1.5}$$

$$H_o = \left(h_o - h - \frac{\rho \frac{\partial h}{\partial T}\Big|_P}{\frac{\partial \rho}{\partial T}\Big|_T} \right) \tag{1.6}$$

$$\tau = V \left(\rho \frac{\partial h}{\partial P}\Big|_T - \frac{\rho \frac{\partial \rho}{\partial P}\Big|_T \cdot \frac{\partial h}{\partial T}\Big|_P}{\frac{\partial \rho}{\partial T}\Big|_P} - 1 \right) \tag{1.7}$$

$$C = \frac{1}{V \frac{\partial \rho}{\partial T}\Big|_P} \tag{1.8}$$

$$D = \frac{\frac{\partial \rho}{\partial P}\Big|_T}{\frac{\partial \rho}{\partial T}\Big|_P} \tag{1.9}$$

The temperature of the superheater is controlled by the attemperator. Therefore, the input mass flow rate to the superheater is the addition of the SC steam and the water spray from the attemperator. The amount of attemperator water spray is regulated by opening the spray valve which responds to a signal from the PI controller. This prevents the high temperature fluctuation and ensures maximum efficiency over a wide range of operation.

1.2.3.2 Fluid Flow

The fluid flow in boiler tubes for one-phase flow is :

$$\dot{m} = k \cdot \sqrt{\Delta P} \tag{1.10}$$

Equation 1.10 is the simplest mathematical expression for fluid flow in boiler tubes. The flow out from the reheater and main steam respectively are:

$$\dot{m}_{rh} = K_1' \frac{P_{rh}}{\sqrt{T_{rh}}} \chi_{rh} \tag{1.11}$$

$$\dot{m}_{ms} = K_2' \frac{P_{ms}}{\sqrt{T_{ms}}} \chi_{ms} \tag{1.12}$$

The detailed derivation of (1.11) and (1.12) can be found in [11].

1.2.4 Turbine/Generator Model

1.2.4.1 Turbine Model

The turbine is modeled through energy balance equations and then is combined with the boiler model.

The work done by high pressure and intermediate pressure turbines are:

$$\dot{W}_{hp} = \dot{m}_{ms} \cdot (h_{ms} - h_{out}) \tag{1.13}$$

$$\dot{W}_{ip} = \dot{m}_{rh} \cdot (h_{rh} - h_{out}) \tag{1.14}$$

The mechanical power of the plant:

$$P_{me} = \dot{W}_{hp} + \dot{W}_{ip} \tag{1.15}$$

Up to Eq. 1.14, the boiler-turbine unit is model in a set of combined equations and can be used for simulation if we assume that the generator is responding instantaneously. However, the dynamics of the turbines' speeds and torques must be affected by the generator dynamics and injecting the mechanical power only into the generator model will not provide this interaction between the variables. To have a strong coupling between the variables in the models of the turbine-generator, torque equilibrium equations for the turbine model are added to the turbine model:

$$\dot{\omega}_{hp} = \frac{1}{M_{hp}}\left[\Gamma_{hp} - D_{hp}\omega_{hp} - K_{HI}(\theta_{hp} - \theta_{ip})\right] \tag{1.16}$$

$$\dot{\theta}_{hp} = \omega_b(\omega_{hp} - 1) = (\omega_{hp} - 1) \tag{1.17}$$

$$\dot{\omega}_{ip} = \frac{1}{M_{ip}}\left[\Gamma_{ip} - D_{ip}\omega_{ip} + K_{HI}(\theta_{hp} - \theta_{ip}) - K_{IG}(\theta_{hp} - \theta_g)\right] \tag{1.18}$$

$$\dot{\theta}_{ip} = \omega_b(\omega_{ip} - 1) = (\omega_{ip} - 1) \tag{1.19}$$

Note that, for two-pole machine: $\theta_g = \delta$

1.2.4.2 Generator Model

The generator models are reported in a number of literatures; a third order nonlinear model is adopted in our work [12]:

$$\dot{\delta} = \Delta\omega \tag{1.20}$$

$$J\Delta\dot{\omega} = \Gamma_a = \Gamma_m - \Gamma_e - D\Delta\omega \tag{1.21}$$

$$\dot{e}'_q = \frac{1}{T'_{do}}\left(E_{FD} - e'_q - \left(x_d - x'_d\right)i_d\right) \tag{1.22}$$

$$\Gamma_e(\text{p.u}) \approx P_e(\text{p.u}) \approx \frac{V}{x'_d}e'_q \sin\delta + \frac{V^2}{2}\left(\frac{1}{x_q} - \frac{1}{x'_d}\right)\sin 2\delta \tag{1.23}$$

1.3 Model Parameter Identification

1.3.1 Identification Procedures

The parameters of the model which are defined by the formulae from (1.3) to (1.7) and the other parameters of mass flow rates' gains, heat transfer constants, turbine, and generator parameters are all identified by Genetic Algorithms in a sequential manner. Even though some of these parameters are inherently not constant, these parameters are fitted directly to the actual plant response to save time and effort. Various data sets of boiler responses have been chosen for identification and verification. First, the parameters of pressure derivatives equations are indentified. Then, the identification is extended to include the temperature equations, the turbine model parameters and finally generator model parameters.

The measured responses which are chosen for identification and verification are:

- Reheater pressure.
- Main SC steam pressure.
- Main SC steam temperature.
- Mass flow rate of SC steam from boiler main outlet to HP turbine.
- Mass flow rate of reheated steam from reheater outlet to the IP turbine.
- Turbine speed.
- Infinite bus frequency.
- Generated power of the plant.

In recent years, Genetic Algorithms optimization tool has been widely used for nonlinear system identification and optimization due to its many advantages over conventional mathematical optimization techniques. It has been proved that the GAs tool is a robust optimization method for parameters identification of sub-critical boiler models [13]. Initially, the GAs produces random values for all the parameters to be identified and called the initial population. Then, it calculates the corresponding fitness function to recopy the best coded parameter in the next generation. The GAs termination criteria depend on the value of the fitness function. If the termination criterion is not met, the GA continues to perform the three main operations which are reproduction, crossover, and mutation. The fitness function for the proposed task is:

$$ff = \sum_{n=1}^{N} (R_m - R_{si})^2 \tag{1.24}$$

The fitness function is the sum of the square of the difference between measured and simulated responses for each of the variables mentioned in this section. N is the number of points of the recorded measured data, The load-up and load-down data have been used for identification. The changes are from 30% to 100% of load and down to 55% to verify the model derived. The model is verified from a ramp load up data and steady state data to cover a large range of once-through operation. The model has been also verified by a third set of data. The GAs parameters setting for identification are listed below:

Generation: 100
Population type: double vector
Creation function: uniform
Population size: 50–100
Mutation rate: 0.1
Mutation function: Gaussian
Migration direction: forward
Selection: stochastic uniform

Figure 1.2 shows some of the load-up identification results. It has been observed that the measured and simulated responses are very well matched for the power generated and they are also reasonably matched for the temperature. Some parameters of the boiler model are listed in Table 1.1 and for heat transfer rates are listed in Table 1.2.

1.3.2 Model Parameter Verification

The validation of the proposed model has been performed using a number of data sets which are the load down and steady state data. Figure 1.3 shows some of the simulated verification results (load-down and steady state simulation). From the results presented, it is obvious that the model response and the actual plant response are well agreed to each other.

1.4 Concluding Remarks

A mathematical model for coal fired power generation with the supercritical boiler has been presented in the paper. The model is based on thermodynamic laws and engineering principles. The model parameters are identified using on-site operating

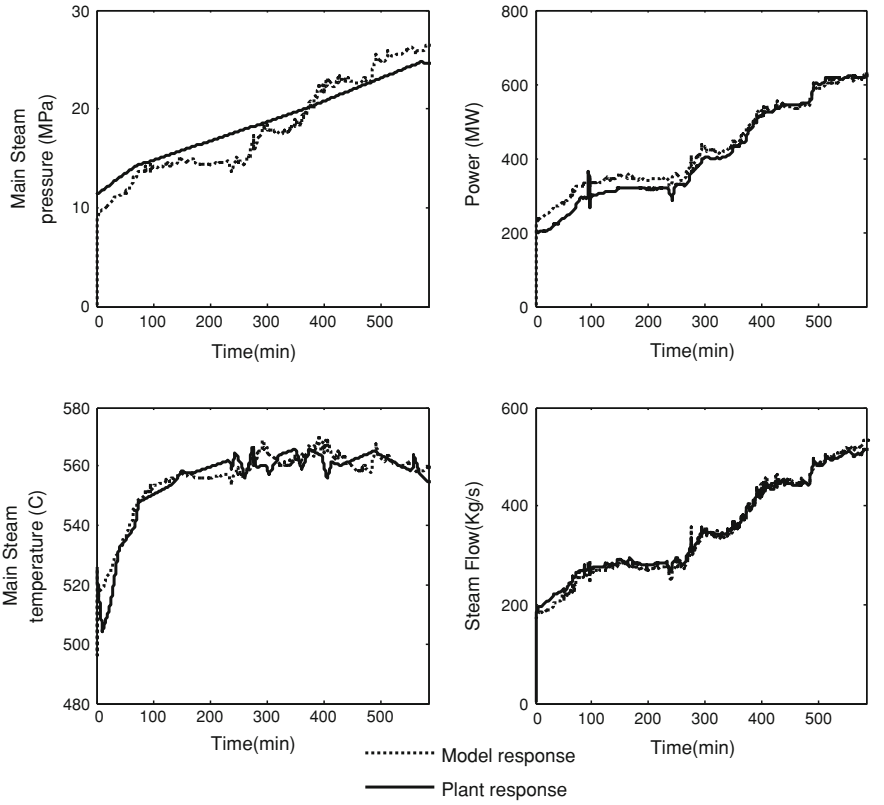


Fig. 1.2 Identification results

Table 1.1 Heat exchanger parameter

HX	H_i	H_o	C	D
ECON	10.2	13.6	2.1e-6	−3.93
WW	12.2	13.3	−1.2e-6	−0.1299
SH	20.5	45.9	1e-6	−3.73
RH	19.8	22.0	−1e-6	−17.9

Table 1.2 Heat transfer rate

τ_1(s)	K_{ec}	K_{ww}	K_{sh}	K_{rh}
9.3	5.7785	7.78	23.776	21.43

data recorded. The model is then verified by using different data sets and the simulation results show a good agreement between the measured and simulated data. For future work, the model will be combined with a nonlinear mathematical

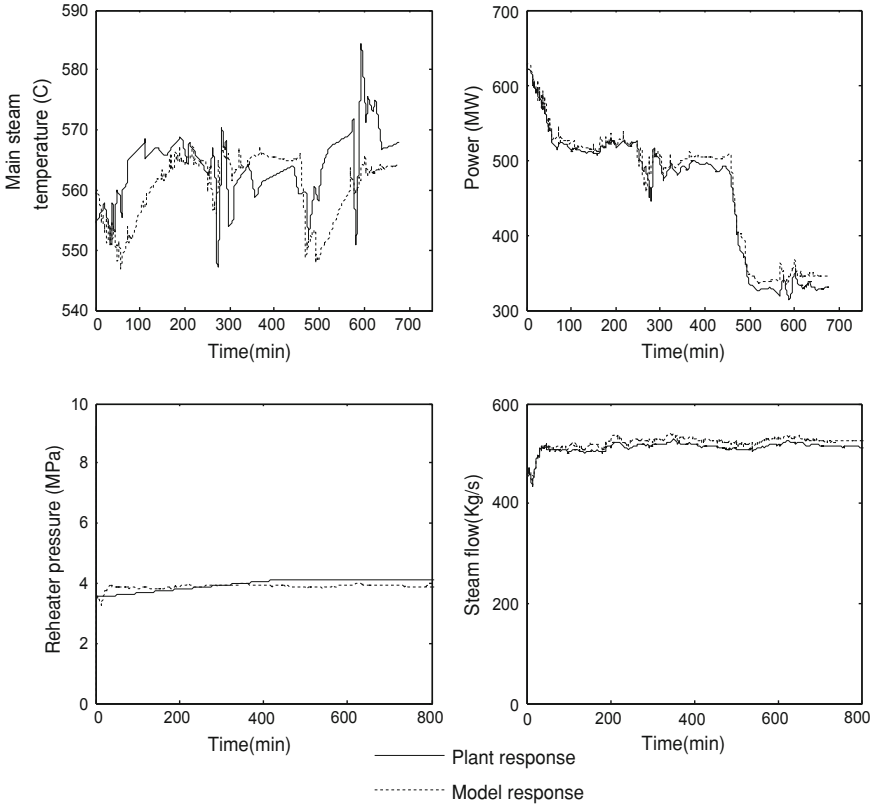


Fig. 1.3 Verification results

model of coal mill to obtain a complete process mathematical model from coal preparation to electricity generation. It is expected that the mill local control system should have great contributions in enhancing the overall control of the plant.

Acknowledgments The authors would like to give our thanks to E.ON Engineering for their support and engineering advices. The authors also want to thank EPSRC (RG/G062889/1) and ERD/AWM Birmingham Science City Energy Efficiency and Demand Reduction project for the research funding support.

References

1. Kundur P (1981) A survey of utility experiences with power plant response during partial load rejection and system disturbances. IEEE Trans Power Apparatus Syst PAS-100(5): 2471–2475
2. Laubli F, Fenton FH (1971) The flexibility of the supercritical boiler as a partner in power system design and operation: part I. IEEE Trans Power Apparatus Syst PAS-90(4): 1719–1724

3. Laubli F, Fenton FH (1971) The flexibility of the supercritical boiler as a partner in power system design and operation: part II. IEEE Trans Power Apparatus Syst PAS-90(4): 1725–1733
4. Littman B, Chen TS (1966) Simulation of bull-run supercritical generation unit. IEEE Trans Power Apparatus Syst 85:711–722
5. Suzuki Y, Sik P, Uchida Y (1979) Simulation of once-through supercritical boiler. Simulation 33:181–193
6. Shinohara W, Kotischek DE (1995) A simplified model based supercritical power plant controller. In: Proceeding of the 35th IEEE Conference on Decision and Control, vol 4, pp 4486–4491
7. Inoue T, Taniguchi H, Ikeguchi Y (2000) A Model of Fossil Fueled Plant with Once-through Boiler for Power System Frequency Simulation Studies. IEEE Trans Power Syst 15(4): 1322–1328
8. Lee KY, Hoe JS, Hoffman JA, Sung HK, Won HJ (2007) Neural network based modeling of large scale power plant. IEEE Power Engineering Society General Meeting No (24–28):1–8
9. Mohamed O, Wang J, Guo S, Al-Duri B, Wei J (2010) Modelling study of supercritical power plant and parameter identification using genetic algorithms. In: Proceedings of the World Congress on Engineering II, pp 973–978
10. Adams J, Clark DR, Luis JR, Spanbaur JP (1965) Mathematical modelling of once-through boiler dynamics. IEEE Trans Power Apparatus Syst 84(4):146–156
11. Salisbury JK (1950) Steam turbines & their cycles. Wiley, New York
12. Yu Y-N (1983) Electric power system dynamics. Academic Press, New York
13. Ghaffari A, Chaibakhsh A (2007) A simulated model for a once through boiler by parameter adjustment based on genetic algorithms. Simul Model Pract Theory 15:1029–1051

Chapter 2
Sequential State Computation Using Discrete Modeling

Dumitru Topan and Lucian Mandache

Abstract In this paper we present a sequential computation method of the state vector, for pre-established time intervals or punctually. Based on discrete circuit models with direct or iterative companion diagrams, the proposed method is intended to a wide range of analog dynamic circuits: linear or nonlinear circuits with or without excess elements or magnetically coupled inductors. Feasibility, accessibility and advantages of applying this method are demonstrated by the enclosed example.

2.1 Introduction

The discretization of the circuit elements, followed by corresponding companion diagrams, leads to discrete circuit models associated to the analyzed analog circuits [1–3]. Using the Euler, trapezoidal or Gear approximations [4, 5], simple discretized models are generated, whose implementation leads to an auxiliary active resistive network. In this manner, the numerical computation of desired dynamic quantities becomes easier and faster. Considering the time constants of the circuit, the discretization time step can be adjusted for reaching the solution optimally, in terms of precision and computation time.

D. Topan (✉)
Faculty of Electrical Engineering, University of Craiova, 13 A.I. Cuza Str., Craiova, 200585, Romania
e-mail: dtopan@central.ucv.ro

L. Mandache
Faculty of Electrical Engineering, University of Craiova, 107 Decebal Blv., Craiova, 200440, Romania
e-mail: lmandache@elth.ucv.ro

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_2,
© Springer Science+Business Media B.V. 2011

The discrete modeling of nonlinear circuits assumes an iterative process too, that requires updating the parameters of the companion diagram at each iteration and each integration time step [5, 6]. If nonzero initial conditions exist, they are computed usually through a steady state analysis performed prior to the transient analysis.

The discrete modeling can be associated to the state variables approach [6, 7], as well as the modified nodal approach [5, 8], the analysis strategy being chosen in accordance with the circuit topology, the number of the energy storage circuit elements (capacitors and inductors) and the global size of the circuit.

The known computation algorithms based on the discrete modeling allow the sequential computation, step by step, along the whole analysis time, of the state vector or output vector directly [5, 9, 10]. In this paper, one proposes a method that allows computing the state vector punctually, at the moments considered significant for the dynamic evolution of the circuit. Thus, the sequential computation for pre-established time subdomains is allowed.

2.2 Modeling Through Companion Diagrams

The time domain analysis is performed for the time interval $[t_0, t_f]$, bounded by the initial moment t_0 and the final moment t_f. It can be discretized with the constant time step h, chosen sufficiently small in order to allow using the Euler, trapezoidal or Gear numerical integration algorithms [1–5]. One can choose $t_0 = 0$ and $t_f = wh$, where w is a positive integer.

The analog circuit analysis using discrete models requires replacing each circuit element through a proper model according to its constitutive equations. In this way, if the Euler approximation is used, the discretization equations and the corresponding discrete circuit models associated to the energy storage circuit elements are shown in Table 2.1, for the time interval $[nh,\ (n+1)h]$, $h<w$.

The tree capacitor voltages $\mathbf{u}_C$ and the cotree inductor currents $\mathbf{i}_L$ [7, 8] are chosen as state quantities, assembled in the state vector $\mathbf{x}$. The currents $\mathbf{I}_C$ of the tree capacitors and the voltages across the cotree inductors $\mathbf{U}_L$ are complementary variables, assembled in the vector $\mathbf{X}$.

At the moment $t = nh$, the above named vectors are partitioned as:

$$\mathbf{x}^n = \begin{bmatrix} \mathbf{u}_C^n \\ \mathbf{i}_L^n \end{bmatrix}, \quad \mathbf{X}^n = \begin{bmatrix} \mathbf{I}_C^n \\ \mathbf{U}_L^n \end{bmatrix} \tag{2.1}$$

with obvious significances of the vectors $\mathbf{u}_C^n,\ \mathbf{i}_L^n,\ \mathbf{I}_C^n,\ \mathbf{U}_L^n$.

For the magnetically coupled inductors, the discretized equations and the companion diagram are shown in Table 2.1, where the following notations were used:

$$\begin{aligned} R_{11}^{n+1} &= \frac{L_{11}}{h}, \quad R_{12}^{n+1} = \frac{L_{12}}{h}, \quad e_1^{n+1} = \frac{L_{11}}{h} i_1^n + \frac{L_{12}}{h} i_2^n, \\ R_{22}^{n+1} &= \frac{L_{22}}{h}, \quad R_{21}^{n+1} = \frac{L_{21}}{h}, \quad e_2^{n+1} = \frac{L_{22}}{h} i_2^n + \frac{L_{21}}{h} i_1^n. \end{aligned} \tag{2.2}$$

Table 2.1 Discrete modeling of the energy storage elements

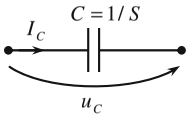
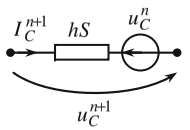
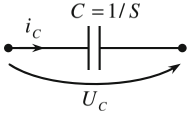
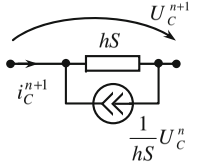
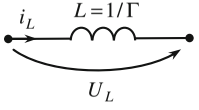
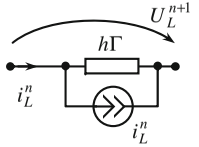
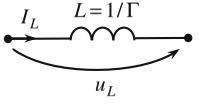
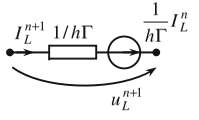
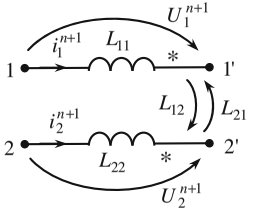
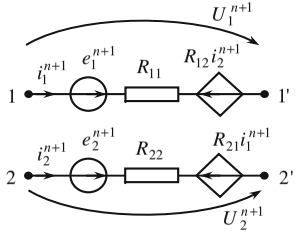
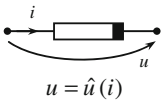
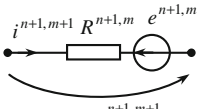
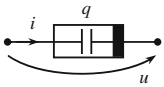
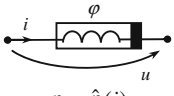
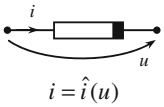
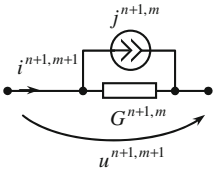
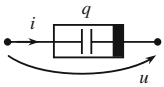
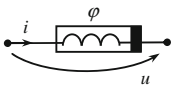
Element	Symbol	Discretized expressions	Companion diagram
Tree capacitor		$u_C^{n+1} = u_C^n + hSI_C^{n+1}$	
Excess capacitor		$i_C^{n+1} = \frac{1}{hS}\left(U_C^{n+1} - U_C^n\right)$	
Cotree inductor		$i_L^{n+1} = i_L^n + h\Gamma U_L^{n+1}$	
Excess inductor		$u_L^{n+1} = \frac{1}{h\Gamma}\left(I_L^{n+1} - I_L^n\right)$	
Magnetically coupled inductor pair		$U_1^{n+1} = R_{11}i_1^{n+1} - R_{11}i_1^n$ $+R_{12}i_2^{n+1} - R_{12}i_2^n$ $U_2^{n+1} = R_{21}i_1^{n+1} - R_{21}i_1^n$ $+R_{22}i_2^{n+1} - R_{22}i_2^n$	

Table 2.2 Iterative discrete modeling

Element	Iterative dynamic parameter	Companion diagram	Notations in the companior diagram
 $u = \hat{u}(i)$	$R^{n+1,m} = \left(\frac{\partial u}{\partial i}\right)_{i=i^{n+1,m}}$		$R^{n+1,m} = R^{n+1,m}$ $e^{n+1,m} = u^{n+1,m} - R^{n+1,m}.$ $i^{n+1,m}$
 $u = \hat{u}(q)$	$C^{n+1,m} = \left(\frac{\partial q}{\partial u}\right)_{u=u^{n+1,m}}$		$R^{n+1,m} = hS^{n+1,m}$ $e^{n+1,m} = u^{n+1,m} - hS^{n+1,m}.$ $i^{n+1,m}$
 $\varphi = \hat{\varphi}(i)$	$L^{n+1,m} = \left(\frac{\partial \varphi}{\partial i}\right)_{i=i^{n+1,m}}$		$R^{n+1,m} = \frac{1}{h}L^{n+1,m}$ $e^{n+1,m} = u^{n+1,m} - \frac{1}{h}L^{n+1,m}.$ $i^{n+1,m}$
 $i = \hat{i}(u)$	$G^{n+1,m} = \left(\frac{\partial i}{\partial u}\right)_{u=u^{n+1,m}}$		$G^{n+1,m} = G^{n+1,m}$ $j^{n+1,m} = i^{n+1,m} - G^{n+1,m}.$ $u^{n+1,m}$
 $q = \hat{q}(u)$	$S^{n+1,m} = \left(\frac{\partial u}{\partial q}\right)_{q=q^{n+1,m}}$		$G^{n+1,m} = \frac{1}{h}C^{n+1,m}$ $j^{n+1,m} = i^{n+1,m} - \frac{1}{h}C^{n+1,m}.$ $u^{n+1,m}$
 $i = \hat{i}(\varphi)$	$\Gamma^{n+1,m} = \left(\frac{\partial i}{\partial \varphi}\right)_{\varphi=\varphi^{n+1,m}}$		$G^{n+1,m} = h\Gamma^{n+1,m}$ $j^{n+1,m} = i^{n+1,m} - h\Gamma^{n+1,m}.$ $u^{n+1,m}$

For nonlinear circuits, the state variable computation at the moment $t = (n+1)h$ requires an iterative process that converges towards the exact solution [4, 5]. A second upper index corresponds to the iteration order (see Table 2.2). Similar results to those of Tables 2.1 and 2.2 can be obtained using the trapezoidal [5, 11] or Gear integration rule [4, 5].

2.3 Sequential and Punctual State Computation

The treatment with discretized models assumes substituting the circuit elements with companion diagrams, which consist in a resistive model diagram. It allows the sequential computation of the circuit solution.

2.3.1 Circuits Without Excess Elements

If the given circuit does not contain capacitor loops nor inductor cutsets [7, 8], the discretization expressions associated to the energy storage elements (Table 2.1, lines 1 and 3), using the notations (2.1), one obtains

$$\mathbf{x}^{n+1} = \mathbf{x}^n + h \begin{bmatrix} \mathbf{S} & 0 \\ 0 & \Gamma \end{bmatrix} \mathbf{X}^{n+1}, \tag{2.3}$$

where $\mathbf{S}$ is the diagonal matrix of capacitor elastances and Γ is the matrix of inductor reciprocal inductances.

Starting from the companion resistive diagram, the complementary variables are obtained as output quantities [5, 10, 11] of the circuit

$$\mathbf{X}^{n+1} = \mathbf{E}\,\mathbf{x}^n + \mathbf{F}\,\mathbf{u}^{n+1}, \tag{2.4}$$

where $\mathbf{E}$ and $\mathbf{F}$ are transmittance matrices, and $\mathbf{u}^{n+1}$ is the vector of input quantities [7, 8] at the moment $t = (n+1)h$.

From (2.3) and (2.4) one obtains an equation that allows computing the state vector sequentially, starting from its initial value $\mathbf{x}^0 = \mathbf{x}(0)$ until the final value $\mathbf{x}^w = \mathbf{x}(wh)$:

$$\mathbf{x}^{n+1} = \mathbf{M}\,\mathbf{x}^n + \mathbf{N}\,\mathbf{u}^{n+1}, \tag{2.5}$$

where

$$\mathbf{M} = 1 + h \begin{bmatrix} \mathbf{S} & 0 \\ 0 & \Gamma \end{bmatrix} \mathbf{E}, \tag{2.6}$$

$\mathbf{1}$ being the identity matrix, and

$$\mathbf{N} = h \begin{bmatrix} \mathbf{S} & 0 \\ 0 & \Gamma \end{bmatrix} \mathbf{F}. \tag{2.7}$$

Starting from Eq. 2.5, through mathematical induction, the useful formula is obtained as

$$\mathbf{x}^n = \mathbf{M}^n \mathbf{x}^0 + \sum_{i=1}^{n} \mathbf{M}^{n-k} \mathbf{N}\,\mathbf{u}^k, \tag{2.8}$$

where the upper indexes of the matrix $\mathbf{M}$ are integer power exponents. The formula (2.8) allows the punctual computation of the state vector at any moment $t = nh$, if the initial conditions of the circuit and the excitation quantities are known.

If a particular solution $\mathbf{x}_p(t)$ of the state equation exists, it significantly simplifies the computation of the general solution $\mathbf{x}(t)$. Using the Euler numerical integration method, one obtains [5]:

$$\mathbf{x}^{n+1} = \mathbf{M}\left(\mathbf{x}^n - \mathbf{x}_p^n\right) + \mathbf{x}_p^{n+1}. \tag{2.9}$$

The sequentially computation of the state vector implies the priory construction of the matrix E, according to Eqs. 2.6 and 2.9. This action requires analyzing an auxiliary circuit obtained by setting all independent sources to zero in the given circuit.

Starting from Eq. 2.9, the expression

$$\mathbf{x}^n = \mathbf{M}^n\left(\mathbf{x}^0 - \mathbf{x}_p^0\right) + \mathbf{x}_p^n \tag{2.10}$$

allows the punctual computation of the state vector.

2.3.2 Circuits with Excess Elements

The excess capacitor voltages [8, 11], assembled in the vector $\mathbf{U}_C$, as well as the excess inductor currents [5, 7, 8], assembled in the vector $\mathbf{I}_L$, can be expressed in terms of the state variables and excitation quantities, at the moment $t = nh$:

$$\begin{bmatrix} \mathbf{U}_C^n \\ \mathbf{I}_L^n \end{bmatrix} = \begin{bmatrix} \mathbf{K}_1 & 0 \\ 0 & \mathbf{K}_2 \end{bmatrix} \mathbf{x}^n + \begin{bmatrix} \mathbf{K}_1' & 0 \\ 0 & \mathbf{K}_2' \end{bmatrix} \mathbf{u}^n, \tag{2.11}$$

where the matrices $\mathbf{K}_1$, $\mathbf{K}_1^{'}$ and $\mathbf{K}_2$, $\mathbf{K}_2^{'}$ contain voltage and current ratios respectively.

Using the Table 2.1, the companion diagram associated to the analyzed circuit can be obtained, whence the complementary quantities are given by:

$$\mathbf{X}^{n+1} = \mathbf{E}\,\mathbf{x}^n + \mathbf{E}_1 \begin{bmatrix} \mathbf{U}_C^n \\ \mathbf{I}_L^n \end{bmatrix} + \mathbf{F}\,\mathbf{u}^n, \tag{2.12}$$

the matrices $\mathbf{E}$, $\mathbf{E}_1$ and $\mathbf{F}$ containing transmittance coefficients.

Considering Eqs. 2.11 and 2.12, the recurrence expression is obtained from (2.5), allowing the sequential computation of the state vector:

$$\mathbf{x}^{n+1} = \mathbf{M}\,\mathbf{x}^n + \mathbf{N}\,\mathbf{u}^{n+1} + \mathbf{N}_1\,\mathbf{u}^n, \tag{2.13}$$

where

$$\begin{aligned} \mathbf{M} &= 1 + h \begin{bmatrix} \mathbf{S} & 0 \\ 0 & \Gamma \end{bmatrix} (\mathbf{E} + \mathbf{E}_1\mathbf{K}), \\ \mathbf{N} &= h \begin{bmatrix} \mathbf{S} & 0 \\ 0 & \Gamma \end{bmatrix} \mathbf{F}, \quad \mathbf{N}_1 = h \begin{bmatrix} \mathbf{S} & 0 \\ 0 & \Gamma \end{bmatrix} \mathbf{E}_1\mathbf{K}', \\ \mathbf{K} &= \begin{bmatrix} \mathbf{K}_1 & 0 \\ 0 & \mathbf{K}_2 \end{bmatrix}, \quad \mathbf{K}' = \begin{bmatrix} \mathbf{K}_1' & 0 \\ 0 & \mathbf{K}_2' \end{bmatrix}. \end{aligned} \tag{2.14}$$

If $\mathbf{x}_p$ is a particular solution of the state equation, the following identity is obtained:

$$\mathbf{N}\,\mathbf{u}^{n+1} + \mathbf{N}_1\mathbf{u}^n = \mathbf{x}_p^{n+1} - \mathbf{M}\,\mathbf{x}_p^n, \tag{2.15}$$

that allows converting (2.13) in the form (2.9), as common expression for any circuit (with or without excess elements).

2.4 Example

In order to exemplify the above described algorithm, let us consider the transient response of the circuit shown in Fig. 2.1, caused by turning on the switch. The circuit parameters are: $R_1 = R_2 = R_3 = 10\,\Omega$; $L = 10\,\text{mH}$; $C = 100\,\mu\text{F}$; $E = 10\text{V}$; $J = 1\text{A}$.

The time-response of capacitor voltage and inductor current will be computed for the time interval $t \in [0,\ 5\,\text{ms}]$. These quantities are the state variables too. The corresponding discretized Euler companion diagram is shown in Fig. 2.2.

According to the notations used in Sect. 2.2, we have:

$$\mathbf{x} = \begin{bmatrix} u_C \\ i_L \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} I_C \\ U_L \end{bmatrix}; \quad \mathbf{u} = \begin{bmatrix} E \\ J \end{bmatrix}$$

The computation way of the matrices $\boldsymbol{E}$ and $\boldsymbol{F}$ arises from the particular form of the expression (2.4):

$$\begin{bmatrix} I_C^{n+1} \\ U_L^{n+1} \end{bmatrix} = \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} \cdot \begin{bmatrix} u_C^n \\ i_L^n \end{bmatrix} + \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix} \cdot \begin{bmatrix} E \\ J \end{bmatrix}$$

from where:

$$e_{11} = \left.\frac{I_C^{n+1}}{u_C^n}\right|_{i_L^n=0;\ E=0;\ J=0}; \quad e_{12} = \left.\frac{I_C^{n+1}}{i_L^n}\right|_{u_C^n=0;\ E=0;\ J=0};$$

$$e_{21} = \left.\frac{U_L^{n+1}}{u_C^n}\right|_{i_L^n=0;\ E=0;\ J=0}; \quad e_{22} = \left.\frac{U_L^{n+1}}{i_L^n}\right|_{u_C^n=0;\ E=0;\ J=0};$$

Using the diagram of Fig. 2.2, the elements of the matrices $\mathbf{E}$ and $\mathbf{F}$ were computed, assuming a constant time step $h = 0.1\,\text{ms}$:

$$\mathbf{E} = \begin{bmatrix} -0.1729 & -0.7519 \\ 0.7519 & -9.7740 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} 0.0827 & 0.8270 \\ 0.0752 & 0.7519 \end{bmatrix}$$

The matrices $\mathbf{M}$ and $\mathbf{N}$ given by Eqs. 2.6, 2.7 are:

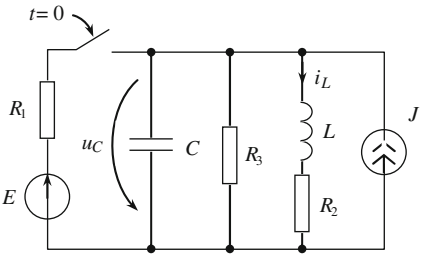


Fig. 2.1 Circuit example

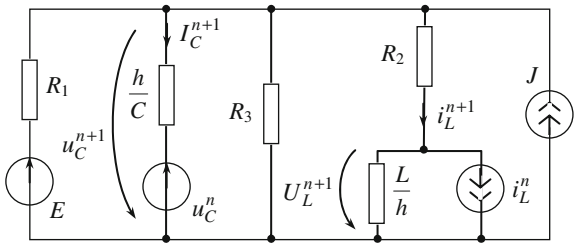


Fig. 2.2 Discretized diagram

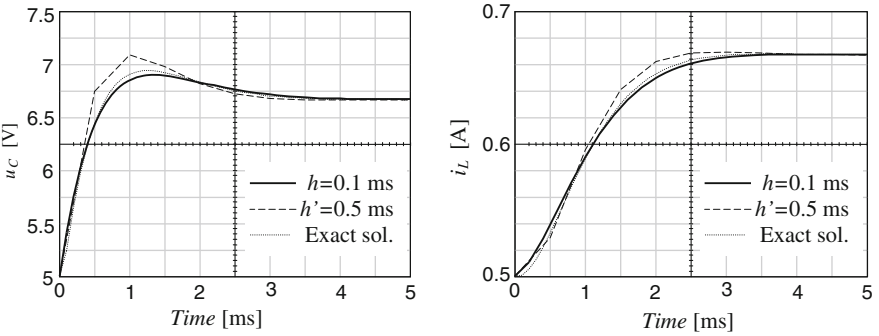


Fig. 2.3 Circuit response

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + 0.1 \cdot 10^{-3} \cdot \begin{bmatrix} \frac{1}{100 \cdot 10^{-6}} & 0 \\ 0 & \frac{1}{10 \cdot 10^{-3}} \end{bmatrix} \cdot \mathbf{E} = \begin{bmatrix} 0.8271 & -0.7519 \\ 0.0075 & 0.9023 \end{bmatrix},$$

$$\mathbf{N} = 0.1 \cdot 10^{-3} \cdot \begin{bmatrix} \frac{1}{100 \cdot 10^{-6}} & 0 \\ 0 & \frac{1}{10 \cdot 10^{-3}} \end{bmatrix} \cdot \mathbf{F} = \begin{bmatrix} 0.0827 & 0.8270 \\ 0.0008 & 0.0075 \end{bmatrix}.$$

Starting from the obvious initial condition

$$\mathbf{x}^0 = \begin{bmatrix} u_C^0 \\ i_L^0 \end{bmatrix} = \begin{bmatrix} 5\ \text{V} \\ 0.5\ \text{A} \end{bmatrix},$$

the solutions were computed using (2.8) and represented in Fig. 2.3 with solid line.

The calculus was repeated in the same manner for a longer time step, $h' = 5h = 0.5\,\text{ms}$, the solution being shown in the same figure. Both computed solutions are referred to the exact solution represented with thin dashed line.

2.5 Conclusion

The proposed analysis strategy and computation formulae allow not only the punctual computation of the state vector, but also allow crossing the integration subdomains with variable time step. The proposed method harmonizes naturally with any procedure based on discrete models of analog circuits, including the methods for iterative computation of nonlinear dynamic networks.

The versatility of the method has already allowed an extension, in connection to the modified nodal approach.

Acknowledgments This work was supported in part by the Romanian Ministry of Education, Research and Innovation under Grant PCE 539/2008.

References

1. Topan D, Mandache L (2010) Punctual state computation using discrete modeling. Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering, vol 2184, London, June 30–July 2 2010, pp 824–828
2. Henderson A (1990) Electrical networks. Edward Arnold, London, pp 319–325
3. Topan D (1978) Computerunterstutze Berechnung von Netzwerken mit zeitdiskretisierten linearisierten Modellen. Wiss. Zeitschr. T.H. Ilmenau, pp 99–107
4. Gear C (1971) The automatic integration of ordinary differential equations. ACM 14(3):314–322
5. Topan D, Mandache L (2007) Chestiuni speciale de analiza circuitelor electrice. Universitaria, Craiova, pp 115–143
6. Topan D (1995) Iterative models of nonlinear circuits. Ann Univ Craiova Electrotech 19:44–48
7. Rohrer RA (1970) Circuit theory: an introduction to the state variable approach. Mc Graw-Hill, New York, pp 3–4
8. Chua LO, Lin PM (1975) Computer-aided analysis of electronic circuits–algorithms and computational techniques. Prentice-Hall, Englewood Cliffs, Chaps. 8–9
9. Chen W-K (1991) Active network analysis. World Scientific, Singapore, pp 465–470
10. Opal A (1996) Sampled data simulation of linear and nonlinear circuits. IEEE Trans Computer-Aided Des Integr Circuits Syst 15(3):295–307
11. Boite R, Neirynck J (1996) Traité d'Electricité, vol IV: Théorie des Réseaux de Kirchhoff. Presses Polytechniques et Universitaires Romandes, Lausanne, pp 146–158

Chapter 3
Detection and Location of Acoustic and Electric Signals from Partial Discharges with an Adaptative Wavelet-Filter Denoising

Jesus Rubio-Serrano, Julio E. Posada and Jose A. Garcia-Souto

Abstract The objective of this research work is the design and implementation of a post-processing algorithm or "search and localization engine" that will be used for the characterization of partial discharges (PD) and the location of the source in order to assess the condition of paper-oil insulation systems. The PD is measured with two acoustic sensors (ultrasonic PZT) and one electric sensor (HF ferrite). The acquired signals are conditioned with an adaptative wavelet-filter which is configured with only one parameter.

3.1 Introduction

The degraded insulation is a main problem of the power equipment. The reliability of power plants can be improved by a preventive maintenance based on the condition assessment of the electrical insulation within the equipments. The insulation is degraded during the period in service due to the accumulation of mechanical, thermal and electric stresses.

Partial discharges (PD) are stochastic electric phenomena that cause a large amount of small shortcoming (<500 pC) inside the insulation [1–3].

J. Rubio-Serrano (✉) · J. E. Posada · J. A. Garcia-Souto
GOTL, Department of Electronic Technology, Carlos III University of Madrid, c/Butarque 15, 28911, Leganes, Madrid, Spain
e-mail: jrserran@ing.uc3m.es

J. E. Posada
e-mail: jposada@ing.uc3m.es

J. A. Garcia-Souto
e-mail: jsouto@ing.uc3m.es

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_3,
© Springer Science+Business Media B.V. 2011

PD are present in the transformers due to the gas dissolved in the oil, the humidity and other faults. They become a problem when PD activity is persistent in time or in a localized area. These are signs of an imminent failure of the power equipment. Thus, the detection, the identification [4] and the localization of PD sources are important tools of diagnosis.

This paper deals with the design of the algorithm that processes the time-series and performs the statistical analysis of the signals acquired in the framework of the MEDEPA test bench in order to assess the insulation faults. This set-up is an experimental PD generation and measurement system designed in the University Carlos III of Madrid in order to study and develop electrical and ultrasonic sensors [5] and analysis techniques, which allow the characterization and the localization of PD.

A PD is an electrical fast transient which produces a localized acoustic emission (AE) due to thermal expansion of the dielectric material [3]. It also generates chemical changes, light emission, etc. [6, 7]. In this work acoustic and electrical signals are processed together. AE is characterized and both methods of detection are put together to assess the activity of PD. The electro-acoustic conversion ratio of PD can be explored by these means [8].

3.2 Experimental Set-Up

The measurements are taken from the MEDEPA experimental set-up. It has the following blocks to generate different types of PD and acquire the signals (up to 100 MSps) from different sensors:

1. *PD generation* the experimental set-up generates controlled PD from a high-voltage AC excitation that is reliable for the ultrasonic sensor characterization and the acoustic measurements.
2. *Instrumentation for electrical measurement* the calibrated electrical measurement allows the correlation of generated PD and provides their basic characteristics (charge, instant of time, etc.).
3. *Instrumentation for acoustic measurements* ultrasonic PZT detectors are used for measuring the AE outside the tank. Fiber-optic sensors are being developed for measurements inside [9].

The experimental set-up is an oil-filled tank with immersed electrodes that generate PD. The ultrasonic sensors (R15i, 150 kHz, $\sim$1 V/Pa) are externally mounted on the tank walls. A wide-band ferrite (10 MHz) is used for electrical measurements and additional instrumentation (Techimp) provides electrical PD analysis. AE travels through the oil (1.5 mm/μs) and the PMMA wall (2.8 mm/μs) to several ultrasonic PZT sensors.

The mechanical and acoustic set-up is represented in Fig. 3.1a. The internal PD generator consists of two cylindrical electrodes of 6 cm of diameter that are separated by several isolating paper layers. High-voltage AC at 50 Hz is applied

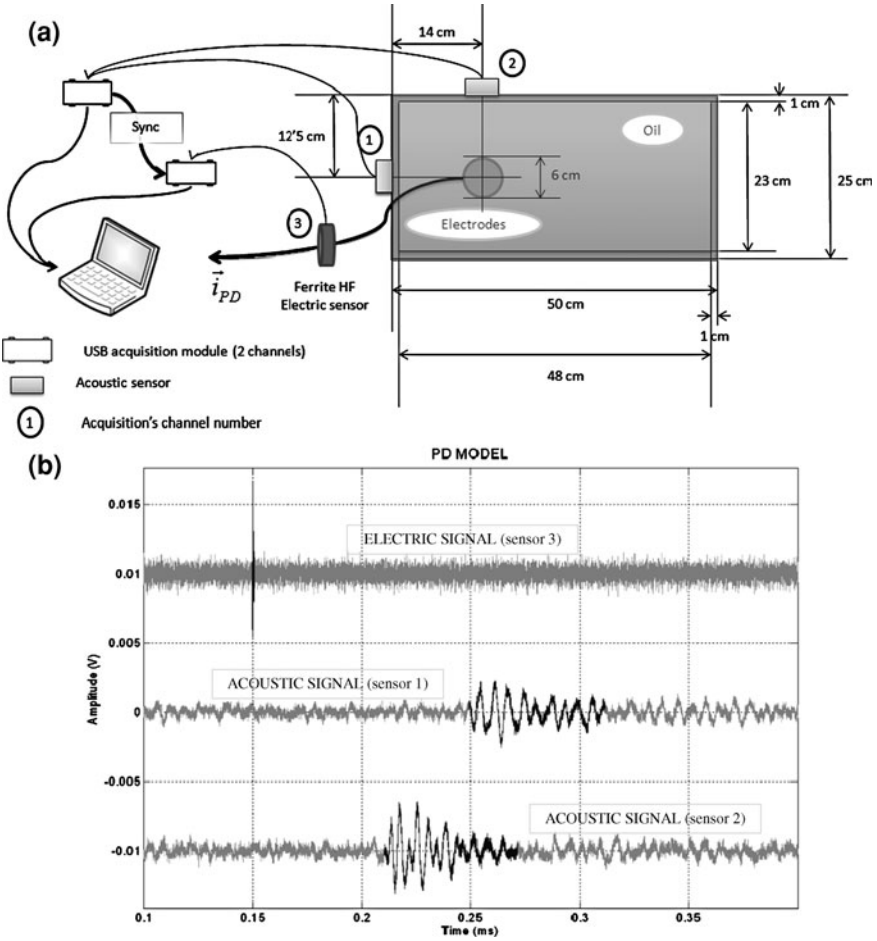


Fig. 3.1 Experimental set-up for acoustic detection and location (**a**). PD single event observation: electric signal and acoustic signals from sensors 1 and 2 (**b**)

between 4.3 and 8.7 kV, so PD are about 100 pC. The expected signals from a PD are as shown in Fig. 3.1b: a single electric signal and an acoustic signal for each channel. The delay is calculated between the electric and acoustic signals to locate spatially the PD source.

Each PD single event produces an electric charge displacement of short-duration (1 μs) that is far shorter than the detected acoustic burst. The electric pulses are detected in the generation circuit. The AE signals are detected in front of the electrodes at the same height on two different walls of the tank. Several sensors are used to obtain the localization of the PD source and the electro-acoustic identification of the PD.

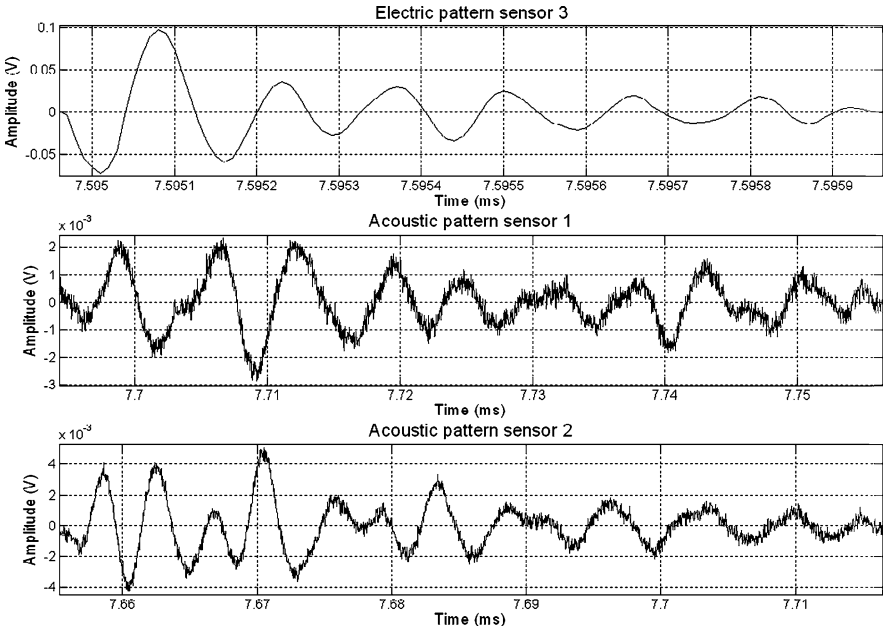


Fig. 3.2 PD electro-acoustic pattern: electric pattern, acoustic patterns of channels 1 and 2

3.3 Signals Characteristics

The detection of PD by electro-acoustic means has the following difficulties: the stochastic process of PD generation and the detection limits of electric and acoustic transients (signal level, identification and matching).

The signals are acquired without any external synchronization due to their stochastic generation. A threshold with the AE signals is setting for assuring at least one AE detection. Afterwards, the time series are analyzed without any reference to the number of PD signals or their time-stamps. AE signals are necessary for the PD spatial location, but they are often less in number than the electric signals due to their strong attenuation caused by the propagation through the oil and the obstacles in the acoustic path.

The AE detection in the experiment has the following characteristics specifically: amplitude usually below 10 mV and signal distortion due to the acoustic propagation path from the PD source to the PZT sensor. In addition, the acoustic angle of incidence to the sensor on the wall produces internal reflection and reverberation. These effects modify the shape, the energy and the power spectrum of the received signal, thus an AE from a single PD is detected differently depending on the position of the sensor. Figure 3.2 shows the characteristic transient waves at each sensor (electric and acoustic) that are associated to a single PD event. This is the electro-acoustic pattern. Though the AE signals are from the same PD their characteristics are different.

Electric signals are easily detected in this experiment. They are used as a zero time reference to calculate the acoustic time of flight from the PD source to the AE sensor. Thus, electro-acoustic processing is performed on the base of pairing the signals from different sensors and sensor types.

Electric and AE signals show diverse duration: 1 μs (electric), 100 μs (AE). Multiple PD from the same or different sources can be generated in the time duration of an AE signal, so the detected AE signal can be the result of the acoustic interference of several PD events. In addition, each AE signal can be associated with more than one electric signal by using time criteria. First approach deals with a processing of the different signals independently and the statistical analysis to link them together and identify PD events [10]. In addition, an all-acoustic system of four or more channels is ongoing to locate PD events upon the basis of a multi-channel processing.

3.4 Signal Processing

The main objective of the algorithm is to analyze the time series in order to detect and evaluate statistically the PD activity and its characteristics: PMCC and energy of PD events, energy ratio between channels and delays.

The selected processing techniques meet the following requirements [10]: (a) same processing regardless the characteristics of the signal, (b) accurate time-stamp of the detected signals, (c) detection based on the shape and the energy, (d) identification tools and (e) statistical analysis for signals pairing.

The signal processing is done with the following structure: pattern selection, wavelet filtering, acoustic detection, electro-acoustic pairing, PD event identification and PD localization.

3.4.1 Pattern Selection

A model of PD is selected form the measurement of a single event (Fig. 3.1b). It is selected by one of these means: (a) technician's observation of a set of signals repetitively with an expected delay, (b) the set of transients selected by amplitude criteria in each channel and (c) a previously stored PD that is useful to study the aging of the insulation. The PD pattern is the set of selected transient waves (Fig. 3.2).

3.4.2 Wavelet Filtering

Signal denoising is performed by wavelet filtering that preserves the time and shape characteristics of the original PD signals [11, 12]. In addition, wavelet filters

are self-configurable for different kind of signals by using an automatic selection rule that extract the main characteristics [13, 14].

The basic steps of the wavelet-filtering are the following: (a) transformation of the signal into the wavelet space, (b) thresholding of the wavelet components (all coefficients smaller than a certain threshold are set to zero) and (c) reverse transformation of the non-zero components. As a result the signal is obtained without undesired noise.

The wavelet-transform is considered two-dimensional: in time and in scale or level of the wavelet. Each level is associated to particular frequency bands. After the n-level transformation the signal in the wavelet-space is a sum of wavelet decompositions (D) and approximations (A):

$$signal = A_n + \sum_{i=1}^{n} D_i \tag{3.1}$$

This tool is used in combination with the Pearson product-moment correlation coefficient (PMCC or ratio) and the energy from the cross-correlation to identify which indices of Di have the main information of the pattern. The same indices are used to configure the filter that is applied to the acquired signals. PMCC is a statistical index that measures the linear dependence between two vectors X and Y. It is independent of the signal's energy so it is used to compare the wave shape of two signals with the same length, although they were out of phase. It is defined by (3.2).

$$PMCC = r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{3.2}$$

The flowchart of the wavelet filtering is shown in Fig. 3.3. It is remarkable that the pattern for each channel is processed only one time. Afterwards the reconstructed signal (PATTERNw) and the configuration of the filter are obtained. Each and every acquisition is configured with these parameters. The filtered patterns (PATTERNw) and the filtered signals (SIGNALw) are the sum of their respective selected decompositions. First the pattern of each channel is filtered and conditioned and then each and every acquisition is individually processed.

This wavelet filter has the following advantages:

- It does not distort the waveform of the signals, so the temporal information is conserved. This is important for cross-correlation and PMCC.
- It does not delay the signal. It is important for time of flight calculation.
- It is self-configurable. Once the threshold is setting, the algorithm selects the decompositions that have the main frequencies of the signals.

3.4.3 Acoustic Detection

Each acoustic acquisition is compared with the acoustic pattern through the cross-correlation. Cross-correlation is used as a measure of the similarity between two signals. Moreover, the time location of each local peak matches with the starting

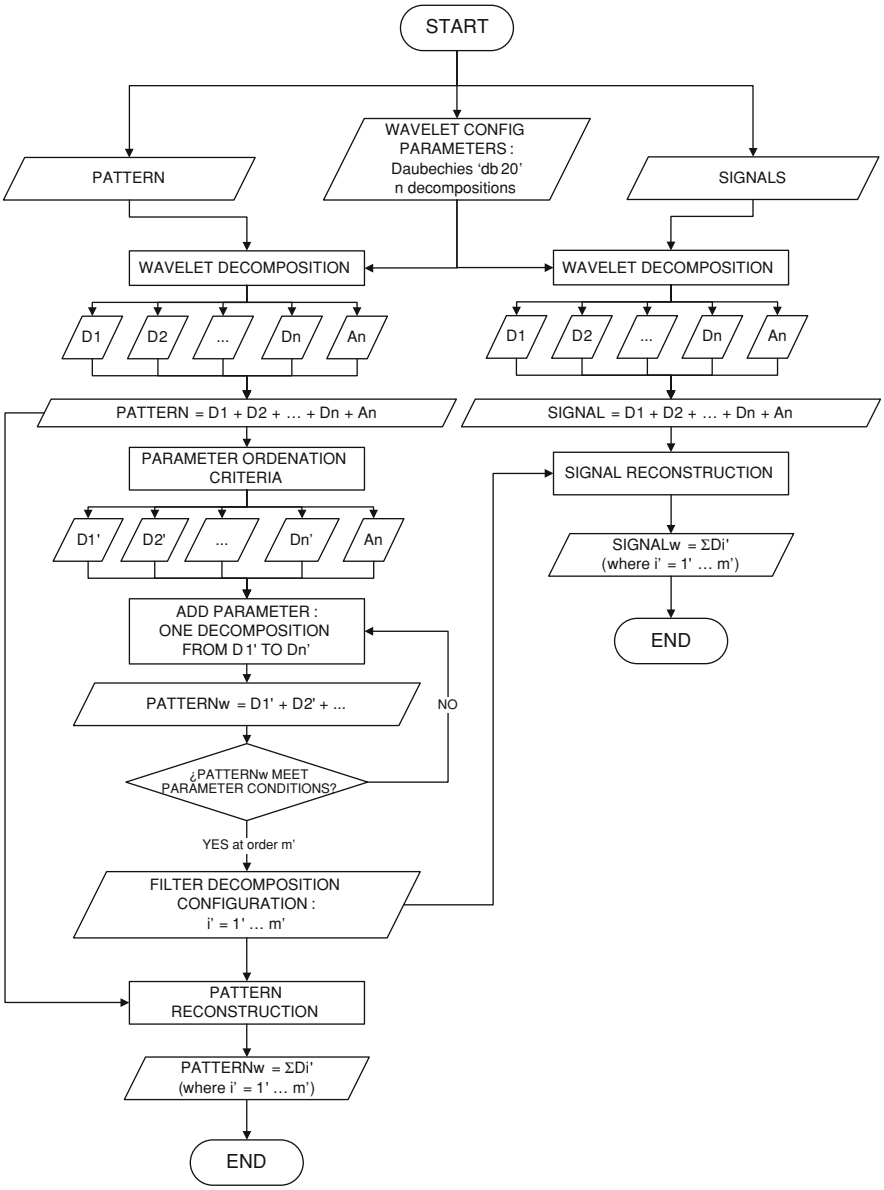


Fig. 3.3 Wavelet filtering flowchart

instant of a transient similar to the pattern. The value of the peak is also a good estimator of the similarity of the signals.

Cross-correlation is used as a search engine to detect the transients that are the best candidates of coming from a PD. It is also used to associate the time-stamp to each one.

The algorithm analyzes the peaks of the cross-correlation in order to decide if the detected transient satisfies the minimum requirements of the selected parameters (energy, amplitude, PMCC, etc.). A maximum of four transients per acoustic acquisition are stored for statistical analysis.

3.4.4 Electro-Acoustic Signal Association

Next step is the cross-correlation of the electric signals with the pattern. Though it is based on the same tool, some differences are introduced. In this case, the maximum absolute value of the cross-correlation is searched. Positive and negative peaks of the cross-correlation are detected and they are associated to the instantaneous phase of the power line voltage, which is an additional parameter for identification.

The electric signals are searched in a temporal window that is compatible with the detected acoustic signal (3.3). Thus, the search within the electric acquisition is delimited between the time-stamp of the acoustic signal and a time period before. This temporal window corresponds with the time that the AE takes to cross the tank. In the experiment of Fig. 3.1 the length of the temporal window is 350 μs by considering $\sim$1.5 mm/μs of sound-speed in oil and 500 mm of the length of the tank.

$$t_{start}(elec\_sig) \in \left(t_{start}(acous\_sig) - \frac{dist.\text{tank}}{v_{sound}},\ t_{start}(acous\_sig) \right) \tag{3.3}$$

Each acoustic signal is matched with up to four electrical signals that satisfy Eq. 3.3 and the database of PD parameters is obtained. Afterwards, the presentation tool provides the histogram of the delay between paired signals in order to analyze the persistency of some delay values. These values with higher incidence correspond to a fault in the insulation. This process of acoustic detection and electro-acoustic signal association is implemented separately for each acoustic channel. The data obtained for each acoustic channel is independent from the others in this approach.

3.4.5 PD Event Association

The association of the transients detected in the acoustic and electric channels provides sets of related signals that come from single PD events with certain probability. Hence, each PD event is defined with three signals: the electric signals that are associated to both acoustic channels and the corresponding acoustic signals. As a result, each PD event contains an electric time-stamp that is the zero time reference and the time of flight of each acoustic signal. These parameters and the references of association are stored in a database as structured information that is used for the statistical analysis.

3.4.6 Localization of the PD Events

Once the database is generated all the PD events are analyzed in order to assess the condition of the insulation. The fault inside the insulation is identified by the persistency of PD events and located acoustically.

The localization is made in the plane which contains the acoustic sensors and the paper between electrodes. PD are generated in this region. Reduced to this 2-D case, Eq. 3.4 is used as a simple localization tool.

$$\begin{aligned}(x_{PD} - x_{S1})^2 + (y_{PD} - y_{S1})^2 &= (v_{sound} \cdot T_{S1})^2 \\ (x_{PD} - x_{S2})^2 + (y_{PD} - y_{S2})^2 &= (v_{sound} \cdot T_{S2})^2\end{aligned} \tag{3.4}$$

Where (x_{S1}, y_{S1}) y (x_{S2}, y_{S2}) are the coordinates of sensors 1 and 2, respectively, and T_{S1} and T_{S2} are the time of flight of the acoustic signals from sensors 1 and 2, respectively. Equation 3.4 represents the intersection of two circumferences whose centers are located in the position of sensors 1 and 2.

When all the PD events are localized and represented, the cluster of PD from the same region is statistically studied in order to find the parameters dependence between acoustic and electric measurements. These PD were probably generated in the same insulation fault so their acoustic path, attenuation and other variables involved in the acoustic detection should be identical.

The PD events of the same cluster are analyzed against the lonely PD events. This study delimitates the range of values for a valid PD event.

The persistency and the concentration of PD activity are the symptoms of the degradation of the insulation system. Hence, thought lonely PD events can be valid, they are no relevant for the detection of faults inside the insulation.

3.5 Experimental Results

The proposed algorithm was applied to process the acquisitions that were taken on MEPEPA test-bench (Fig. 3.1a). In this experiment 76 series of acoustic and electric signals were acquired simultaneously. Each time series is approximately 8 ms and it is sampled at 100 MSps.

First, the electric and acoustic patterns were selected from an isolated PD event (Fig. 3.1b) and they are filtered with the wavelet processing (Fig. 3.3).

The electric pattern is a fast transient of about 7 MHz and its duration is 1 μs approximately. The length of the acoustic pattern is 35 μs and its central frequency is 150 kHz. A detail of the signals involved in the wavelet filtering to obtain the acoustic pattern of one sensor is shown in Fig. 3.4.

A limitation of the selected pattern is the reverberation of the acoustic waves that is detected through the wall. For normal incidence of the acoustic signal on the PMMA wall (sound velocity of 2.8 mm/μs) the reflection takes 7 μs to reach the

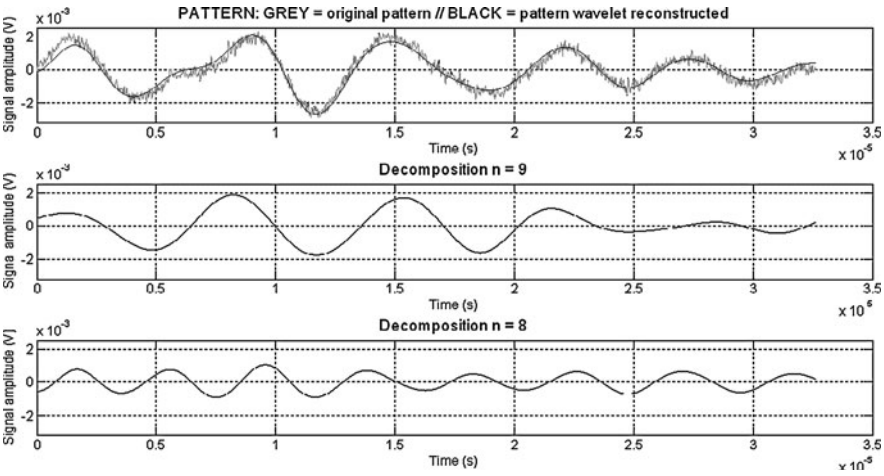

PATTERN: GREY = original pattern // BLACK = pattern wavelet reconstructed
Signal amplitude (V)
Time (s)
Decomposition n = 9
Decomposition n = 8

Fig. 3.4 Acoustic pattern of sensor 1 after reconstruction and its decomposition

detector again (20 mm). Thus the distortion of the acoustic signal is observed from 7 μs onwards.

Once the patterns are selected and filtered each and every acquired signal series is filtered, it is processed with the cross correlation and analyzed with the PMCC. As a result the local peaks of the cross-correlation give the time-stamps that can be associated to PD events. These events are also characterized by their indexes and the transient waveforms that were found in the time series (Fig. 3.5).

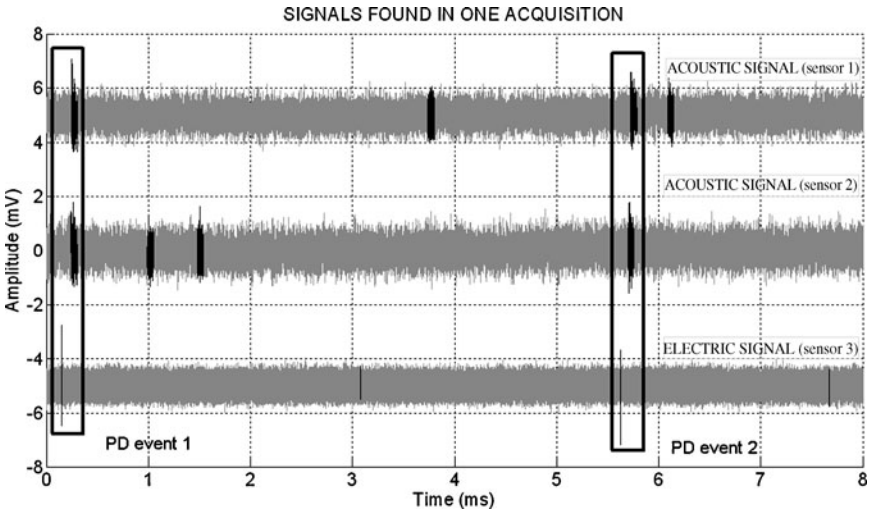
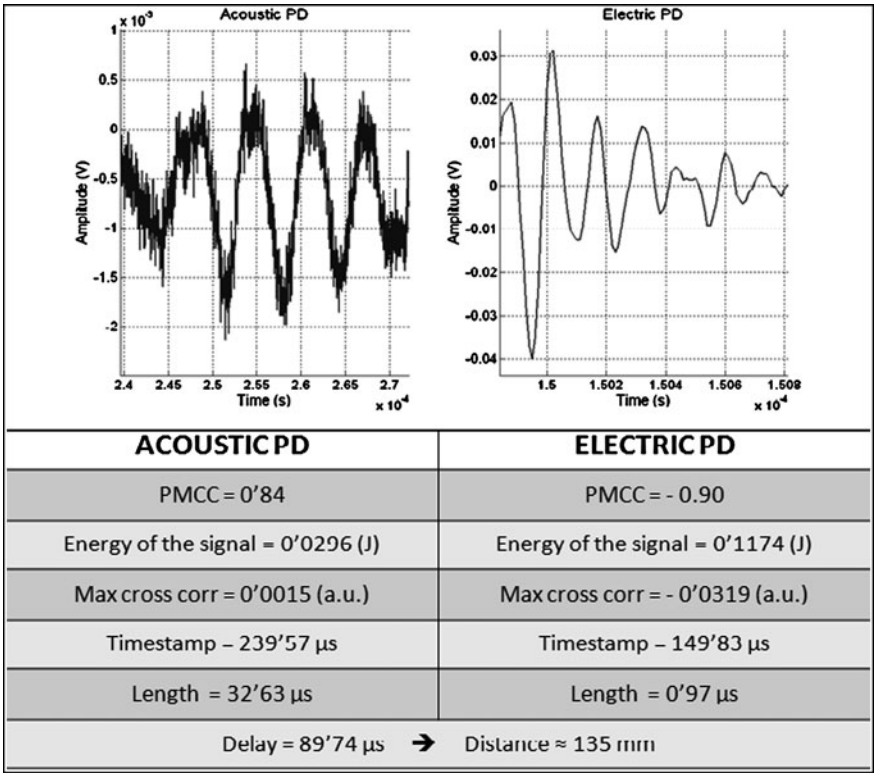


Fig. 3.5 Electric and acoustic time-series. Sets of transient signals found as probable PD events



ACOUSTIC PD	ELECTRIC PD
PMCC = 0'84	PMCC = - 0.90
Energy of the signal = 0'0296 (J)	Energy of the signal = 0'1174 (J)
Max cross corr = 0'0015 (a.u.)	Max cross corr = - 0'0319 (a.u.)
Timestamp – 239'57 μs	Timestamp – 149'83 μs
Length = 32'63 μs	Length = 0'97 μs
Delay = 89'74 μs → Distance ≈ 135 mm	

Fig. 3.6 Example of detected PD event (details of AE signal in sensor 1 and electric signal)

Acoustic and electric signals are matched together and the parameters of each signal are calculated. In Fig. 3.6 there is an example of one of the paired-signals (acoustic sensor 1 and electric sensor 3) found by the algorithm. The parameters of each transient signal and of the pair are also shown.

Now, signals can be classified by their delays. In the experiment, there are some valid values with an incidence of four or more. The delay of maximum incidence is 102 μs (Fig. 3.7) for sensor 1 and 62 μs for sensor 2.

In future work, it will be examined the relation between energy and the PMCC as a function of their location. The goal is to find and discriminate PD not only for its location but also for its expected values of energy and PMCC.

Finally, Eq. 3.4 is employed to locate the origin of the acoustic signals in the plane (Fig. 3.8). It is important to emphasize that lonely PD events are observed and they can be valid events. However, they are no relevant for the detection of faults inside the insulation because their low persistency and their location are not characteristic for the insulation condition assessment.

Figure 3.8 shows the existence of a region inside the electrodes where PD were frequently generated. This region represents a damaged area in the insulation. The

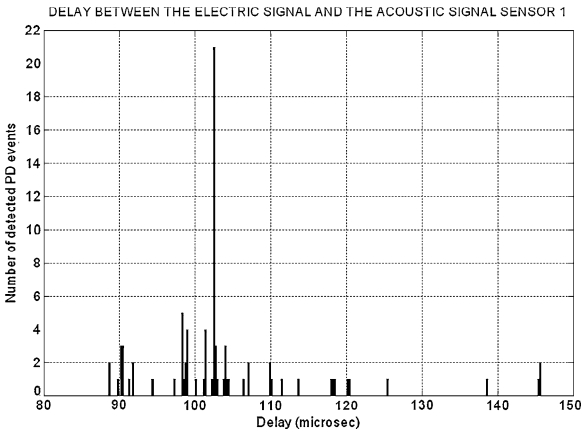


Fig. 3.7 Histogram of delays obtained from the acoustic sensor 1 with the selected PD

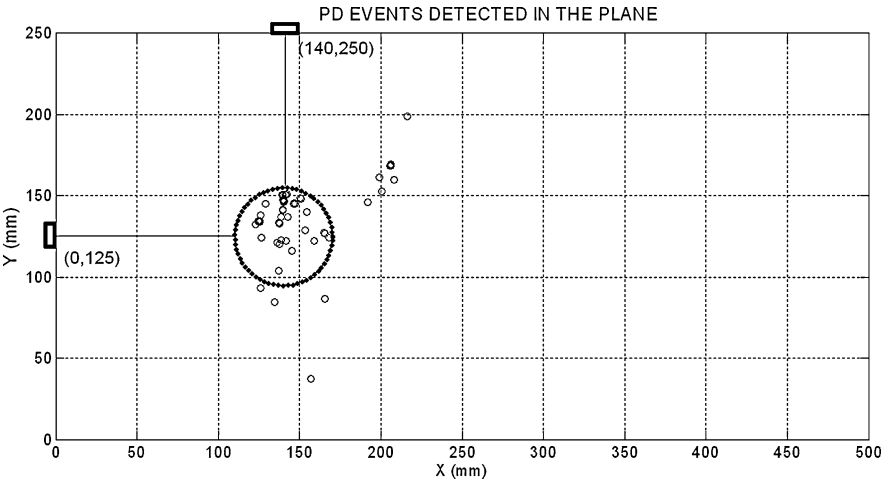


Fig. 3.8 Location of the detected PD events in the plane

concentration of PD in this region is a symptom of an imminent failure of the electrical system.

3.6 Conclusions and Future Work

The design and implementation of a post-processing algorithm is presented for the detection and location of partial discharges and the condition assessment of the insulation. The algorithm is able to parameterize the signals, to define the ranges and to delimitate the time windows in order to locate and classify PD in transformers. It was applied to signals from internal PD that were acquired by external

acoustic sensors, but it is being extended to superficial PD and internal acoustic fiber-optic sensors. The purpose of this signal processing within the framework of the MEDEPA test-bed is to locate, identify and parameterize PD activity to predict imminent failures in insulation systems.

The main features of the proposed algorithm are the following: its feasibility to detect and identify PD signals from different sensors, the adaptability of the wavelet filtering based on an external pattern, and the multi-sensor statistical analysis instead of a single event approach. In addition, the wavelet denoising does not alter the temporal characteristics.

Although the algorithm is not still designed for real-time use, after the temporal series are processed, their parameters are stored in a database, which is used as a reference of PD activity for further studies and extended to the maintenance of transformers in service.

In order to improve the signal detection and identification, the next step is the calibration of the tool with different kind of PD activities (types, intensities and sources) and the statistical analysis of the parameters in the database. In addition to the time windowing and the pattern-matching, other parameters will be considered to assess the probability of PD: persistency, 3-D location, energy and PMCC.

The location of PD sources is of main concern in the application of AE. The objective is to implement a 3-D algorithm compatible with the designed tools, either with external sensors, or using also internal sensors [9]. An all-acoustic system of four or more channels is ongoing to locate PD events upon the basis of a multi-channel processing.

Finally, the electro-acoustic conversion ratio of PD activity is an open research line with the implemented statistical analysis.

Acknowledgments This work was supported by the Spanish Ministry of Science and Innovation, under the R&D projects No. DPI2006-15625-C03-01 and DPI2009-14628-C03-01 and the Research grant No. BES-2007-17322. PD tests have been made in collaboration with the High Voltage Research and Tests Laboratory of Universidad Carlos III de Madrid (LINEALT).

References

1. Bartnikas R (2002) Partial discharges: their mechanism, detection and measurement. IEEE Trans Dielectr Electr Insul 9(5):763–808
2. Van Brunt RJ (1991) Stochastic properties of partial discharges phenomena. IEEE Trans Electr Insul 26(5):902–948
3. Lundgaard LE (1992) Partial discharge—part XIV: acoustic partial discharge detection—practical application. IEEE Electr Insul Mag 8(5):34–43
4. Suresh SDR, Usa S (2010) Cluster classification of partial discharges in oil-impregnated paper insulation. Adv Electr Comput Eng J 10(5):90–93
5. Macia-Sanahuja C, Lamela H, Rubio J, Gallego D, Posada JE, Garcia-Souto JA (2008) Acoustic detection of partial discharges with an optical fiber interferometric sensor. In: IMEKO TC 2 Symposium on Photonics in Measurements
6. IEEE Guide for the Detection and Location of Acoustic Emissions from Partial Discharges in Oil-Immersed Power Transformers and Reactors (2007). IEEE Power Engineering Society

7. Santosh Kumar A, Gupta RP, Udayakumar K, Venkatasami A (2008) Online partial discharge detection and location techniques for condition monitoring of power transformers: a review. In: International Conference on Condition Monitoring and Diagnosis, Beijing, China, 21–24 April 2008
8. von Glahn P, Stricklett KL, Van Brunt RJ, Cheim LAV (1996) Correlations between electrical and acoustic detection of partial discharge in liquids and implications for continuous data recording. Electr Insul 1:69–74
9. Garcia-Souto JA, Posada JE, Rubio-Serrano J (2010) All-fiber intrinsic sensor of partial discharge acoustic emission with electronic resonance at 150 kHz. SPIE Proc 7726:7
10. Rubio-Serrano J, Posada JE, Garcia-Souto JA (2010) Digital signal processing for the detection and location of acoustic and electric signals from partial discharges. In: Proceedings of the World Congress on Engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, vol 2184, issue 1, pp 967–972
11. Ma X, Zhou C, Kemp IJ (2002) Automated wavelet selection and thresholding for PD detection. IEEE Electr Insul Mag 18(2):37–45
12. Keppel G, Zedeck S (1989) Data analysis for research designs—analysis of variance and multiple regression/correlation approaches. Freeman, New York
13. Wang K-C (2009) Wavelet-based speech enhancement using time-frequency adaptation. EURASIP J Adv Signal Process 2009:8
14. Pinle Q, Yan L, Ming C (2008) Empirical mode decomposition method based on wavelet with translation invariance. EURASIP J Adv Signal Process 2008:6

Chapter 4
Study on a Wind Turbine in Hybrid Connection with a Energy Storage System

Hao Sun, Jihong Wang, Shen Guo and Xing Luo

Abstract Wind energy has been focused as an inexhaustible and abundant energy source for electrical power generation and its penetration level has increased dramatically worldwide in recent years. However, its intermittence nature is still a universally faced challenge. As a possible solution, energy storage technology hybrid with renewable power generation process is considered as one of options in recent years. The paper aims to study and compare two feasible energy storage means—compressed air (CAES) and electrochemical energy storage (ECES) for wind power generation applications. A novel CAES structure in hybrid connection with a small power scale wind turbine is proposed. The mathematical model for the hybrid wind turbine system is developed and the simulation study of system dynamics is given. Also, a pneumatic power compensation control strategy is reported to achieve acceptable power output quality and smooth mechanical connection transition.

H. Sun · J. Wang (✉) · S. Guo · X. Luo
School of Electronic, Electrical and Computer Engineering, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
e-mail: j.h.wang@bham.ac.uk

H. Sun
e-mail: hxs823@bham.ac.uk

S. Guo
e-mail: s.guo@bham.ac.uk

X. Luo
e-mail: xxl738@bham.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_4,
© Springer Science+Business Media B.V. 2011

4.1 Introduction

Nowadays, the world is facing the challenge to meet the continuously increasing energy demand and to reduce the harmful impact to our environment. In particular, wind energy appears as a preferable solution to take a considerable portion of the generation market, especially in the UK. However, the key challenge faced by wind power generations is intermittency. The variability of wind power can lead to changes in power output from hour to hour, which arises from changes in wind speed. Figure 4.1 shows that the power output from a diversified wind power system is usually changing hourly from 5 to 20%, either higher or lower [1]. Besides, energy regulatory policies all around the world have been characterized by introducing competition in the power industry and market, both at the wholesale and at the retail levels. The variable market brought uncertain variations onto power transmission and distribution networks, which have been studied at length [2, 3]. It is highly desired to alleviate such impacts through alternative technologies.

One proposed solution is to introduce an element of storage or an alternative supply for use when the ambient flux is insufficient for a guaranteed supply to the demand. The primary cause is that energy storage can make wind power available when it is most demanded. Apart from the pumped water, battery, hydrogen and super-capacitors, compressed air energy storage (CAES) is also a well known controllable and affordable technology of energy storage [4–6]. In a CAES system, the excess power is used to compress air which can be stored in a vessel or a cavern. The energy stored in compressed air can be used to generate electricity when required. Compared with other types of energy storage schemes, CAES is sustainable and will not produce any chemical waste. In this paper, a comparative analysis between CAES and electrochemical energy storage (ECES) has been conducted. A hybrid energy storage wind turbine system is proposed in the paper, which connects a typical wind turbine and vane-type air motor for compressed air energy conversion. The mathematical model for the whole system is derived and simulation study is conducted. The study of such a CAES system has shown a promising merit provided by the proposed hybrid connection of wind turbines and CAES.

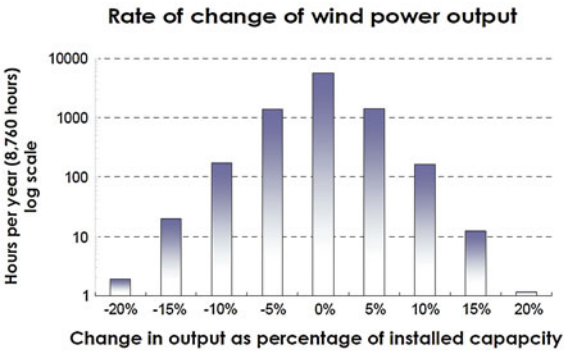


Fig. 4.1 The hourly change of wind power output

4.2 Electrochemical and Compressed Air Energy Storage

In this paper, the feasibility of energy storage for 2 kW household small scale wind turbine is analyzed. Electrochemical energy storage is the most popular type of energy storage in the world from small to large scales. For instance, the lead-acid battery is the oldest rechargeable battery with widest range of applications, which is a mature and cost-effective choice among all the electrochemical batteries. The main advantages of ECES are no emission, simple operation and higher energy efficiency. The efficiency of lead-acid batteries is generally around 80%. While, the compressed air energy storage is also cleaner as no chemical disposal pollution is produced to environment [7]. However, CAES has rather lower energy efficiency; much energy is lost during the process of thermal energy conversion [8–10].

A drawback ECES faced is relatively short lifetime that mainly expressed on the limited charge/discharge cycle life. For example, lead-acid batteries' cycle life is roughly in the range of 500–1500. This issue can be more serious when it is applied to wind power generation due to the high variation in wind speed and low predictability to the wind power variation patterns, that is, the battery will be frequently charged and discharged. For CAES, the pneumatic actuators, including compressor, air motor, tank, pipes and valves, are relatively robust; the major components have up to 50-year lifetime. Therefore, the whole system lift time would be only determined by the majority of the mechanical components in the system.

The capacity of an electrochemical battery is directly related to the active material in the battery. That means the more energy the battery can offer, the more active material will be contained in the battery, and therefore the size, weight as well as the price is almost linear versus the battery capacity. For the compressed air system, the capacity correlates to the volume of the air storage tank. Even though the pneumatic system also requires large space to sustain a long term operation, but it has been proven more cost-effective in consideration of the practically free raw material (see Table 4.1 [11]).

The electromotive force of a lead-acid cell provides only about two Volts voltage due to its electrochemical characteristics, and enormous amount of cells therefore should be connected in series to obtain a higher terminal voltage. With this series connection, if one cell within the battery system goes wrong, the whole battery may fail to store or offer energy in the manner desired. Discouragingly, it is very hard currently to diagnose which cell in the system fails and it is expensive and not cost-effective to replace the whole pack of batteries. Besides, most lead-acid batteries designed for the deep discharge are not sealed, and the regular maintenance is therefore required due to the gas emission caused by the water

Table 4.1 Typical marginal energy storage costs

Types	Overall cost
Electro-chemical storage	>$400/KWh
Pumped storage	$80/KWh
Compressed air	$1/KWh

Table 4.2 Comparison between CAES and ECES

	CAES	ECES
Service life	Long	Short
Efficiency	Not high	Very high
Size	Large depend on tank size	Large depend on cell number
Overall cost	Very cheap	Very expensive
Maintenance	Need regular maintenance	Hard to overhaul, need regular maintenance

electrolysis while overcharged. Comparing with these characteristics of batteries, CAES only needs regular leakage test and oil maintenance. In brief, a comparison between CAES and ECES can be summarized in Table 4.2.

4.3 The Hybrid Wind Turbine System with CAES

There are two possible system structures for a hybrid wind turbine system with compressed air energy storage; one has been demonstrated as an economically solution for utility-scale energy storage on the hours' timescale. The energy storage system diagram is illustrated in Fig. 4.2.

Such systems are successfully implemented in Hantorf in Germany, McIntosh in Alabama, Norton in Ohio, a municipality in Iowa, in Japan and under construction in Israel [12]. The CAES produces power by storing energy in the form of compressed air in an underground cavern. Air is compressed during off-peak periods, and is used on compensating the variation of the demand during the peak periods to generate power with a turbo-generator/gas turbine system. However, this system seems to be disadvantageous as it needs a large space to store compressed air, such as large underground carven for large scale power facilities. So this may limit its applications in terms of site installation. Besides all the above mentioned issues, large-capacity converter and inverter systems are neither cost effective nor power effective.

For smaller capacity of wind turbines, this paper presents a novel hybrid technology to engage energy storage to wind power generation. As shown in Fig. 4.3, the electrical and pneumatic parts are connected through a mechanical transmission mechanism. This electromechanical integration offers simplicity of design, therefore, to ensure a higher efficiency and price quality. Also, the direct compensation of torque variation of the wind turbine will alleviate the stress imposed onto the wind turbine mechanical parts.

4.4 Modelling Study of the Hybrid Wind Turbine System

For the proposed system illustrated in Fig. 4.3, the detailed mathematical model has been derived, which is used to have an initial test for the practicability of the whole hybrid system concept. At this stage, the system is designed to include a

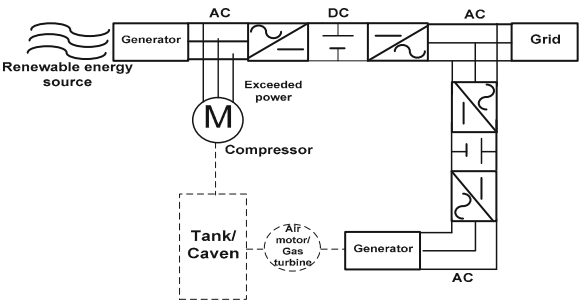


Fig. 4.2 Utility-scale CAES application's diagram

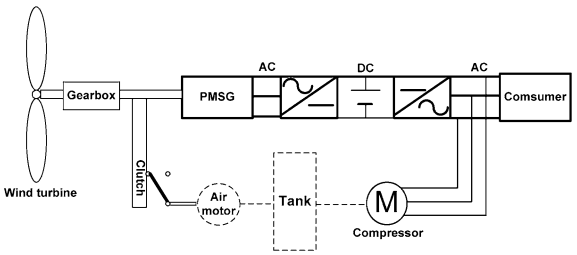


Fig. 4.3 Small scale hybrid wind turbine with CAES

typical wind turbine with a permanent magnetic synchronous generator (PMSG), a vane type air motor and the associated mechanical power transmission system. The pneumatic system can be triggered to drive the turbine for power compensation during the low wind power period. The whole system mathematical model is developed and described below.

4.4.1 Mathematical Model of the Wind Turbine

For a horizontal axis wind turbine, the mechanical power output P that can be produced by the turbine at the steady state is given by:

$$P = \frac{1}{2}\rho\pi r_T^2 v_w^3 C_p \tag{4.1}$$

where ρ is the air density, v_w is the wind speed, r_T is the blade radius; C_p reveals the capability of turbine for converting energy from wind. This coefficient depends on the tip speed ratio $\lambda = \omega_T r_T / v_w$ and the blade angle, ω_T denotes the turbine speed. As this requires knowledge of aerodynamics and the computations are rather complicated, numerical approximations have been developed [13, 14]. Here the following function will be used,

$$C_p(\lambda, \theta) = 0.22\left(\frac{116}{\lambda_i} - 0.4\theta - 5\right)e^{\frac{-12.5}{\lambda_i}} \tag{4.2}$$

with

$$\frac{1}{\lambda_i} = \frac{1}{\lambda + 0.08\theta} - \frac{0.035}{\theta^3 + 1} \tag{4.3}$$

To describe the impact of the dynamic behaviors of the wind turbine, a simplified drive train model is considered.

$$\frac{d}{dt}\omega_T = \frac{1}{J_T}(T_T - T_L - B\omega_T) \tag{4.4}$$

Where J_T is the inertia of turbine blades, T_T and T_L mean the torque of turbine and low speed shaft respectively, B is the damping coefficient of the driven train system.

4.4.2 Modeling the Permanent Magnetic Synchronous Generator (PMSG)

The model of a PMSG with pure resistance load (for simplicity of analysis) is formed of the following equations.

For the mechanical part:

$$\frac{d}{dt}\omega_G = \frac{1}{J_G}(T_G - T_e - F\omega_G) \tag{4.5}$$

$$\frac{d\theta_G}{dt} = \omega_G \tag{4.6}$$

For the electrical part:

$$\frac{d}{dt}i_d = \frac{1}{L_d}v_d - \frac{R_s}{L_d}i_d + \frac{L_q}{L_d}p\omega_G i_q \tag{4.7}$$

$$\frac{d}{dt}i_q = \frac{1}{L_q}v_q - \frac{R_s}{L_q}i_q - \frac{L_d}{L_q}p\omega_G i_d - \frac{\varepsilon p\omega_G}{L_q} \tag{4.8}$$

$$T_e = 1.5p\left[\varepsilon i_q + (L_d - L_q)i_d i_q\right] \tag{4.9}$$

$$V_q = \frac{1}{3}\left[\sin(p\theta_G)\cdot(2V_{ab} + V_{bc}) + \sqrt{3}V_{bc}\cos(p\theta_G)\right] \tag{4.10}$$

$$V_d = \frac{1}{3}\left[\cos(p\theta_G)\cdot(-2V_{ab} - V_{bc}) - \sqrt{3}V_{bc}\sin(p\theta_G)\right] \tag{4.11}$$

where, θ_G and ω_G are generator rotating angle and speed, F means the combined viscous friction of rotor and load, i is current, v means voltage, L is inductance, R_s is resistance of stator windings, p is the number of pole pairs of the generator,

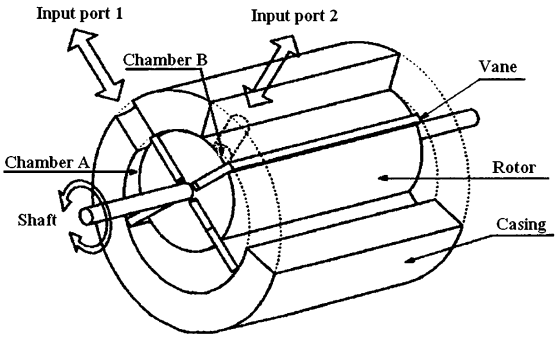


Fig. 4.4 Structure of a vane type air motor with four vanes

ε is the amplitude of the flux induced by the permanent magnets of the rotor in the stator phases. While the subscripts a, b, c, d, q represent the axes of a, b, c, d, q for different electrical phases, respectively. The three-phase coordinates and d–q rotating frame coordinates can be transformed each other through Park's transformation [15].

4.4.3 Model of the Vane-Type Air Motor

Figure 4.4 shows the sketch of a vane-type air motor with four vanes. In this paper, input port 1 is supposed to be the inlet port, and then input port 1 will be outlet port. Compressed air is admitted through the input port 1 from servo valves and fills the cavity between the vanes, housing and rotor. The chamber A which is open to the input port 1 fills up under high pressure. Once the port is closed by the moving vane, the air expands to a lower pressure in a higher volume between the vane and the preceding vane, at which point the air is released via the input port 2. The difference in air pressure acting on the vane results in a torque acting on the rotor shaft [16, 17].

A simplified vane motor structure is shown in Fig. 4.5. The vane working radius measured from the rotor centre x_a can be derived by:

$$x_a = e\cos\phi + \sqrt{R_m^2 - e^2\sin^2\phi} \tag{4.12}$$

The volumes of chamber A and chamber B are derived as follows, and presented by the subscription a and b in this part equations.

$$V_a = \frac{1}{2}L\left(R_m^2 - r^2\right)(\pi + \phi) + \frac{1}{4}L_m e^2\sin 2\varphi + L_m e R_m \sin\varphi \tag{4.13}$$

$$V_b = \frac{1}{2}L_m\left(R_m^2 - r^2\right)(\pi - \phi) - \frac{1}{4}L_m e^2\sin 2\varphi - L_m e R_m \sin\varphi\pi \tag{4.14}$$

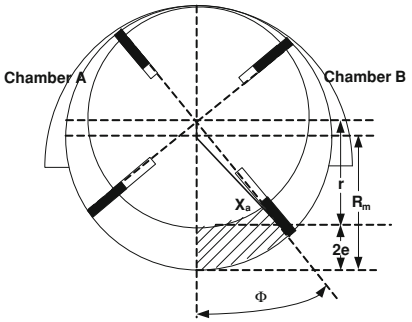


Fig. 4.5 Schematic diagram of the structure of a vane-type air motor

where, R_m is radius of motor body; e is eccentricity; L_m is vane active length in the axial direction, ϕ is motor rotating angle, r means rotor radius.

The pressure of chamber A and B can be derived [10]:

$$\dot{P}_a = -\frac{k\dot{V}_a}{V_a}P_a + \frac{k}{V_a}RT_sC_dC_0A_aX_af(P_a, P_s, P_e) \tag{4.15}$$

$$\dot{P}_b = -\frac{k\dot{V}_b}{V_b}P_b + \frac{k}{V_b}RT_sC_dC_0A_bX_bf(P_b, P_s, P_e) \tag{4.16}$$

where, T_s is supply temperature, R, C_d, C_0 are air constant, A is effective port width of control valve, X is valve spool displacement, f is a function of the ratio between the downstream and upstream pressures at the orifice.

The drive torque is determined by the difference of the torque acting on the vane between the drive and exhaust chambers, and is given by [13]:

$$M = (P_a - P_b)\left(e^2\cos 2\varphi + 2eR_m\cos\varphi + R_m^2 - r^2\right)L/2 \tag{4.17}$$

4.4.4 Model of Mechanical Power Transmission

The power transmission system, which is similar to a vehicle air conditioning system, includes the clutch and the belt speed transmission to ensure coaxial running, as shown in Fig. 4.6 [18].

The clutch will be engaged only when the turbine and air motor operate at the same speed to avoid mechanical damage to the system components. Even so, the system design still faces another challenge during the engagement, that is, the speed of air motor could not reach the speed as high as the turbine generator does, in most instances. Therefore, the two plates of belt transmission are designed in different diameters to play the function as a gearbox does. The main issue of modeling the power transmission is that two different configurations are presented:

Case I Clutch disengaged: After the air motor started during the period before the two sides of electromagnetic clutch get the same speed, the clutch can be

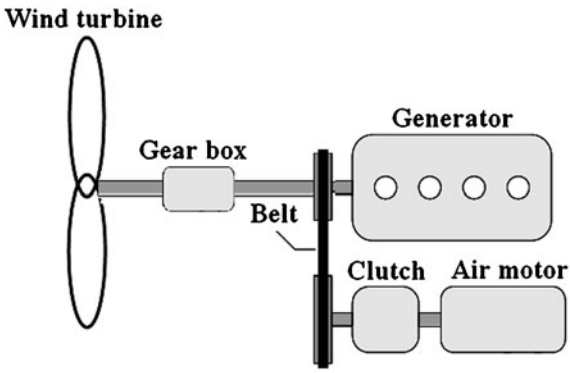


Fig. 4.6 The structure of the power transmission system in hybrid wind turbine

considered completely separated. While the scroll air motor is at the idle status with the inertia load of clutch friction plate. Considering friction and different payloads and applying Newton's second law of angular motion, we have

$$M - M_f\dot{\phi} = (J_a + J_f)\ddot{\phi} \tag{4.18}$$

where J_a is the air motor inertia, J_f is friction plate inertia, M is the drive torque, M_f is the friction coefficient, $\dot{\phi}$ is the angular velocity, $\ddot{\phi}$ represents the angular acceleration.

Both the active plate and passive plate of the belt transmission can be considered as the generator inertia load, so the total equivalent inertia is

$$J_{total} = J_{pass} + \kappa^2 J_{act} \tag{4.19}$$

where J_{pass} and J_{act} is the inertia of passive and active plate respectively, and κ is the speed ratio of the belt.

Case II Clutch engaged: Once the angular velocity of the air motor $\dot{\phi}$ meets the speed of the active plate ω_G/i, the clutch will be engaged with the two sides. After the engagement, the active plate and friction plate can be assumed together to be one mass. The dynamic equations are as follows:

$$\begin{cases} M - M_f\dot{\phi} - T_{act} = (J_a + J_f + J_{act})\ddot{\phi} \\ T_{pass} = \dfrac{T_{act}\eta}{\kappa} \\ \dfrac{d\omega_G}{dt} = \dfrac{1}{J_G + J_{pass}}(T_H + T_{pass} - T_e - F\omega_G) \\ \dot{\phi} = \dfrac{\omega_G}{\kappa} \end{cases}$$

where, T_H is the input torque of wind turbine high speed shaft, η is the transfer efficiency of the belt.

Choose system state variables to be x_1: pressure in the chamber A, x_2: pressure in the chamber B, x_3: rotated angle, x_4: angular speed, x_5: current in d axis,

x_6: current in q axis. And input variables u_1: wind speed, u_2: input valve displacement. Combining the wind turbine, driven train and generator models together, the state functions of the whole hybrid wind turbine system can then be described by:

$$\dot{x}_1 = -\frac{k\dot{V}_a}{V_a}x_1 + \frac{k}{V_a}RT_sC_dC_0A_au_2f(P_a, P_s, P_e)$$

$$\dot{x}_2 = -\frac{k\dot{V}_b}{V_b}x_2 + \frac{k}{V_b}RT_sC_dC_0A_bX_bf(P_b, P_s, P_e)$$

$$\dot{x}_3 = \frac{x_4}{\kappa}$$

$$\dot{x}_4 = \frac{1}{J_G + J_{pass} + J_T\frac{\eta'}{\kappa'^2} + (J_a + J_f + J_{act})\frac{\eta}{\kappa^2}}\left\{\eta\frac{\rho\pi r^2u_1^3C_p'}{2x_5} - \eta\frac{B'x_4}{\kappa'^2} + \eta\frac{M}{\kappa} - \eta\frac{M_fx_4}{\kappa^2} - M_cS\left(\frac{x_4}{\kappa}\right) - \frac{3}{2}p(\varepsilon x_6 + L_dx_6x_5 - L_qx_6x_5) - Fx_4\right\}$$

$$\dot{x}_5 = \frac{v_d}{L_d} - \frac{R_s}{L_d}x_5 + \frac{L_q}{L_d}px_4x_6$$

$$\dot{x}_6 = \frac{v_q}{L_q} - \frac{R_s}{L_q}x_6 - \frac{L_d}{L_q}px_4x_5 - \frac{\varepsilon px_5}{L_q}$$

where, η', κ' is the efficiency and speed ratio of wind turbine gearbox. With such a complicated structure of the system model, sometimes, it is difficult to obtain accurate values of system parameters. Intelligent optimization and identification methods have been proved to be an effective method to tackle this challenging problem [19, 20]. The test system for the proposed hybrid system structure is under development in the authors' laboratory and the data obtained from the rig can be used to improve the model accuracy.

4.5 Simulation Study

The model derived above for the proposed hybrid wind turbine system is implemented in MATLAB/SIMULINK environment to observe the dynamic behavior of the whole system as shown in Fig. 4.7. The simulation results are described below.

The simulation considers the scenario when the input wind speed steps down within a 40 s' time series observation window, that is, drops from 9 to 8 m/s at the time of 20 s and the whole simulation time period is 40 s (see Fig. 4.8).

For comparison, the results from hybrid system using 6 bar supply pressure and those from stand-alone system without pneumatic actuators are shown in Fig. 4.9. It can be seen that the hybrid system can still obtain a high turbine speed due to the contribution of air motor output. It can also maintain a steady value even the natural wind speed decreases. Regrettably however, the power coefficient of

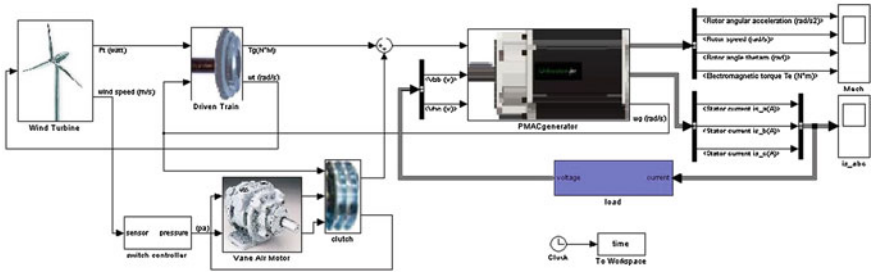


Fig. 4.7 The block diagram of the simulation system

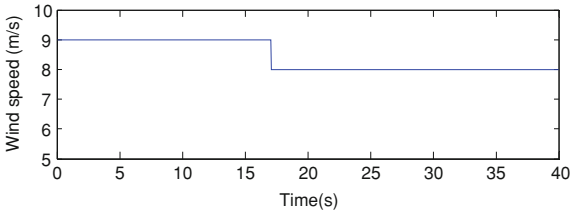
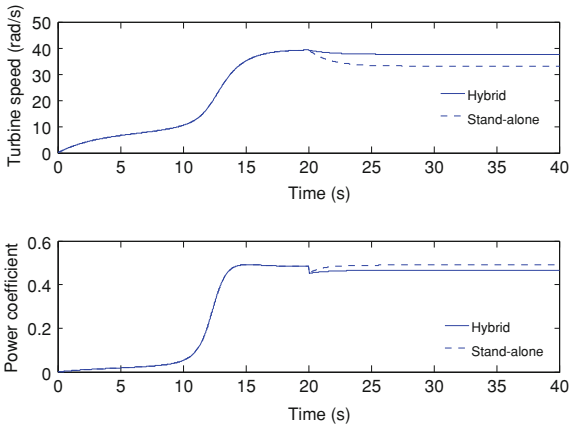


Fig. 4.8 Input wind speed

Fig. 4.9 Simulation results of wind turbine



turbine falls because of the increased tip speed ratio $\lambda = \omega_T r_T / v_w$. That should be considered as adverse effect of the hybrid system.

Figure 4.10 provides a significant contrast between hybrid and independent status through generator operation. It can be seen that the power compensation can almost overcome the energy shortfall at the lower wind speed.

Figure 4.11 reveals the simulation results of vane type air motor. The air motor started at the time of 20 s, and joined the wind turbine system rapidly owing to its fast response characteristic. It is worth noting that this type of air motor should

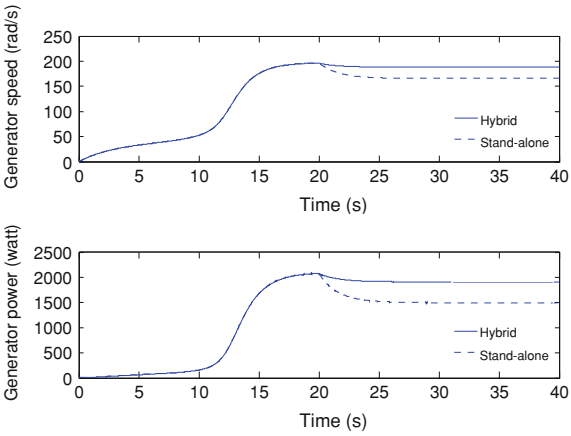


Fig. 4.10 Simulation results of the responses of the PMSG

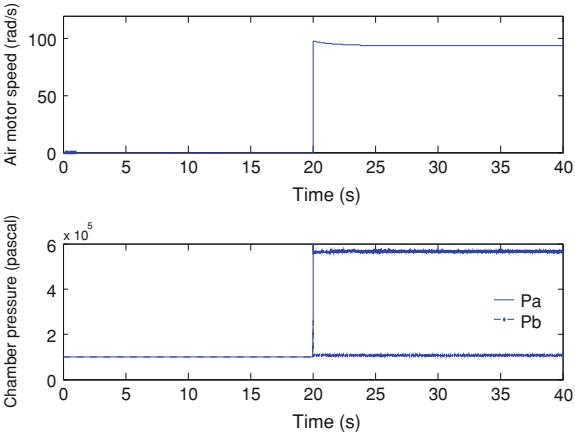


Fig. 4.11 Simulation results of vane type air motor

generally running with well-marked periodic fluctuation, which is originated from the cyclically changed difference between P_a and P_b (the pressures in chamber A and chamber B). However, in hybrid system, the air motor operates rather smoothly which may be resulted from the large inertia of the whole system.

4.6 Concluding Remarks

This paper presents a concise review on two types of energy storage technologies. A new concept of CAES applied to a small power scale wind turbine system is introduced. The complete process mathematical model is derived and implemented under MATLAB/SIMULINK environment. The simulation results are very encouraging as the extra power from the air motor output compensates the power shortfall from wind energy. This strategy enables the wind turbine to operate at a

relatively uniformly distributed speed profile, which in turn will improve the operation condition of the overall system. The simple structure of the system and the advantage of CAES would provide the opportunities for such a system to be placed in the future renewable energy electricity market. The research in hybrid wind turbines is still on-going and further improvement is expected. Advanced tracking control strategy is a promising methodology and currently in consideration by the research team [7, 21].

Acknowledgments The authors would like to thank the support from ERDA/AWM for the support of Birmingham Science City Energy Efficiency & Demand Reduction project, China 863 Project (2009AA05Z212) and the scholarships for Hao Sun, Xing Luo from the University of Birmingham, UK.

References

1. Sinden G (2005) Wind power and UK wind resource. Environmental change institute, University of Oxford
2. Akhmatov V (2002) Variable-speed wind turbines with doubly-fed induction generators. Part II: power system stability. Wind Eng 26(3):71–88
3. Hansena AD, Michalke G (2007) Fault ride-through capability of DFIG wind turbines. Renew Energy 32:1594–1610
4. Cavallo A (2007) Controllable and affordable utility-scale electricity from intermittent wind resources and compressed air energy storage (CAES). Energy 32:120–127
5. Lemofouet S, Rufer A (2006) A hybrid energy storage system based on compressed air and supercapacitors with maximum efficiency point tracking (MEPT). IEEE Trans Ind Electron 53(4):1105–1115
6. Van der Linden S (2006) Bulk energy storage potential in the USA, current developments and future prospects. Energy 31:3446–3457
7. Wang J, Pu J, Moore P (1999) Accurate position control of servo pneumatic actuator systems: an application to food packaging. Control Eng Pract 7(6):699–706
8. Yang L, Wang J, Lu N et al (2007) Energy efficiency analysis of a scroll-type air motor based on a simplified mathematical model. In: The Proceedings of the World Congress on Engineering. London, pp 759–764, 2–4 July 2007
9. Wang J, Yang L, Luo X, Mangan S, Derby JW (2010) Mathematical modelling study of scroll air motors and energy efficiency analysis—Part I. IEEE-ASME Transactions on Mechatronics. doi:10.1109/TMECH.2009.2036608
10. Wang J, Luo X, Yang L, Shpanin L, Jia N, Mangan S, Derby JW (2010) Mathematical modelling study of scroll air motors and energy efficiency analysis—Part II. IEEE-ASME Transactions on Mechatronics. doi:10.1109/TMECH.2009.2036607
11. Price A (2009) The current status of electrical energy storage systems. ESA London Meeting, London, UK
12. Vongmanee V (2009) The renewable energy applications for uninterruptible power supply based on compressed air energy storage system. In: IEEE Symposium on Industrial Electronics and Applications (ISIEA 2009). Kuala Lumpur, Malaysia, 4–6 October 2009
13. Heier S (1998) Grid integration of wind energy conversion systems. Wiley, Chicheste
14. Sun H, Wang J, Guo S, Luo X (2010) Study on energy storage hybrid wind power generation systems. In: The Proceedings of the World Congress on Engineering 2010 WCE 2010, Vol II, London, UK, June 30–July 2, pp 833–838

15. Pillay P, Krishnan R (1989) Modeling, simulation and analysis of permanent magnet motor drives, part 1: the permanent-magnet synchronous motor drive. IEEE Trans Ind Appl 25:265–273
16. Luo X, Wang J, Shpanin L, Jia N, Liu G, Zinober A (2008) Development of a mathematical model for vane-type air motors with arbitrary N vanes. In: International Conference of Applied and Engineering Mathematics, WCE, vol I–II. London, pp 362–367, July 2008
17. Wang J, Pu J, Moore PR, Zhang Z (1998) Modelling study and servo-control of air motor systems. Int J Control 71(3):459–476
18. Yeung YPB, Cheng KWE, Chan WW, Lam CY, Choi WF, Ng TW (2009) Automobile hybrid air conditioning technology. In: The Proceedings of the 3rd International Conference on Power Electronics Systems and Applications, p 116
19. Wei JL, Wang J, Wu QH (2007) Development of a multi-segment coal mill model using an evolutionary computation technique. IEEE Trans Energy Convers 22:718–727
20. Zhang YG, Wu QH, Wang J, Oluwanda G, Matts D, Zhou XX (2002) Coal mill modelling by machine learning based on on-site measurement. IEEE Trans Energy Convers 17(4):549–555
21. Wang J, Kotta U, Ke J (2007) Tracking control of nonlinear pneumatic actuator systems using static state feedback linearization of the input-output map. In: Proceedings of the Estonian Academy of Sciences-Physics Mathematics, vol 56, pp 47–66

Chapter 5
SAR Values in a Homogenous Human Head Model

Levent Seyfi and Ercan Yaldız

Abstract The purpose of this chapter is to present how to determine and reduce specific absorption rate (SAR) on mobile phone user. Both experimental measurement technique and a numerical computing method are expressed here. Furthermore, an application on reduction of SAR value induced in human head is carried out with numerical computing. Mobile phone working at 900 MHz frequency shielded with copper is considered in order to furnish reduction of SAR in simulations which are conducted to calculate the maximum SAR values in Matlab programming language using two dimensional (2D) Finite Difference Time Domain (FDTD) method. Calculations are separately made for both 1 g and 10 g. Head model structure is assumed uniform.

5.1 Introduction

Today mobile phone is one of the most widely used electronic equipments. What is more, it has a large number of users regardless of age. For this reason, designing of mobile phones which do not adversely affect human health is of great importance. The mobile phones are used mostly very close to ear as shown in Fig. 5.1. In this case, electromagnetic (EM) wave of mobile phone mainly radiates towards user's head (that is, brain).

L. Seyfi (✉) · E. Yaldız
Department of Electrical and Electronics Engineering, Selçuk University, Konya, Turkey
e-mail: leventseyfi@selcuk.edu.tr

E. Yaldız
e-mail: eyaldiz@selcuk.edu.tr

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_5,
© Springer Science+Business Media B.V. 2011

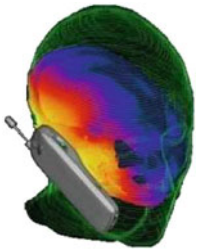


Fig. 5.1 Distribution of EM waves from a mobile phone on human head

Mobile phones communicate by transmitting radio frequency (RF) waves through base stations. RF waves are non-ionizing radiation which cannot break chemical bonds nor cause ionization in the human body. The operating frequencies of mobile phones can change depending on the country and the service provider between 450 and 2700 MHz. The RF radiation to a user mitigates rapidly while increasing distance from mobile phone. Using the phone in areas of good reception decreases exposure as it allows the phone to communicate at low power. A large number of studies have been performed over the last two decades to assess whether mobile phones create a potential health risk [1, 2]. To date, no adverse health effects have been established for mobile phone use [3].

Investigations of effects of mobile phones, other devices emitting EM waves on human health and measures against them have been still continued. As the results were evaluated, 1°C temperature increase of tissue cannot be removed in the circulatory system and this damages tissue. Limits for each frequency band were specified by the relevant institutions according to this criterion. Limits for the general public in the whole body average SAR value and in localized SAR value are 0.08 and 2 W/kg at 10 MHz–10 GHz frequency band, respectively [4]. SAR (W/kg) is the amount of the power absorbed by unit weight tissue. Measuring of SAR values in living cells is not experimentally possible. Specifically created model (phantom) and specialized laboratory test equipment are used for this. SAR values can be measured experimentally by placing probe into the phantom. The equipment consists of a phantom (human or box), precision robot, RF field sensors, and mobile phone holder, as shown in Fig. 5.2. The phantom is filled with a liquid that approximately represents the electrical properties of human tissue.

Determination of SAR values can also be carried out with numerical calculations as an alternative to using the phantom [5–8]. In this case, the calculations are executed with simulations using electrical properties and physical dimensions of the typical human head.

Mobile phones are manufactured within the limited SAR values. However, negative consequences may be seen in time due to placing them close to head during calling and due to long phone calls. In this case, it may be required to use with some precautions. For instance, a headset-microphone set can be used while calling.

Alternatively, the attenuation of EM waves emitted from mobile phone towards user's head by using the conductive material can be provided. Conductive material

Fig. 5.2 Experimentally measuring SAR value with a phantom

mostly reflects the EM waves back. Hence, the amount of absorption of EM waves will be reduced to minimum level by placing the suitable sized conductive plate between the mobile phone's antenna and the user head. To reduce SAR, some studies having different techniques has been introduced, too [9, 10].

In this chapter, 2D-FDTD technique, absorbing boundary conditions, and SAR calculation method are expressed. Additionally, a numerical application is presented. In the application, 2D simulations have been conducted to investigate reducing of SAR values in user head using copper plate. Simulations have been carried out in Matlab programming language using the 2D-FDTD method. First order Mur's boundary condition have been used to remove artificial reflections naturally occurred in FDTD method.

5.2 2D-FDTD Method

When Maxwell's differential equations are considered, it can be seen that the change in the E-field in time is dependent on the change in the H-field across space. This results in the basic FDTD time-stepping relation that, at any point in space, the updated value of the E-field in time is dependent on the stored value of the E-field and the numerical curl of the local distribution of the H-field in space.

Similar situation with above is present for the H-field. Iterating the E-field and H-field updates results in a marching-in-time process wherein sampled-data analogs of the continuous EM waves under consideration propagate in a numerical grid stored in the computer memory. Yee proposed that the vector components of the E-field and H-field spatially stagger about rectangular unit cells of a cartesian computational grid so that each E-field vector component is located midway between a pair of H-field vector components, and conversely [11, 12]. This scheme, now known as a Yee lattice, constructs the core of many FDTD software.

The choices of grid cell size and time step size are very important in applying FDTD. Cell size must be small enough to permit accurate results at the highest operating frequency, and also be large enough to keep computer requirements manageable.

Cell size is directly affected by the materials present. The greater the permittivity or conductivity, the shorter the wavelength at a given frequency and the smaller the cell size required. The cell size must be much less than the smallest wavelength for which accurate results are desired. An often used cell size is $\lambda/10$ or less at the highest frequency. For some situations, such as a very accurate determination of radar scattering cross-sections, $\lambda/20$ or smaller cells may be necessary. On the other hand, good results are obtained with as few as four cells per wavelength. If the cell size is made much smaller than the Nyquist sampling limit, $\lambda = 2\Delta x$, is approached too closely for reasonable results to be obtained and significant aliasing is possible for signal components above the Nyquist limit.

Once the cell size is selected, the maximum time step is determined by the Courant stability condition. Smaller time steps are permissible, but do not generally result in computational accuracy improvements except in special cases. A larger time step results in instability. To understand the basis for the Courant condition, consider a plane wave propagating through an FDTD grid. In one time step, any point on this wave must not pass through more than one cell, because during one time step, FDTD can propagate the wave only from one cell to its nearest neighbors. To determine this time step constraint, a plane wave direction is considered so that the plane wave propagates most rapidly between field point locations. This direction will be perpendicular to the lattice planes of the FDTD grid. For a grid of dimension d (where $d = 1$, 2, or 3), with all cell sides equal to Δu, it is found that with v the maximum velocity of propagation in any medium in the problem, usually the speed of light in free space [13],

$$v\Delta t \leq \frac{\Delta u}{\sqrt{d}} \tag{5.1a}$$

for stability. If the cell sizes are not equal, it is as following for a 2-D and 3-D rectangular grid, respectively [14, 15].

$$v\Delta t \leq 1/\sqrt{\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2}} \tag{5.1b}$$

$$v\Delta t \leq 1/\sqrt{\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2} + \frac{1}{(\Delta z)^2}} \tag{5.1c}$$

where Δt is temporal increment and $\Delta x, \Delta y, \Delta z$, denoting sides of the cubic cell are spatial increments in the x, y, and z-direction, respectively.

Firstly, although the real world is obviously 3D, many useful problems can be solved in two dimensions when one of the dimensions is much longer than the other two. In this case, it is generally assumed that the field solution does not vary in this dimension, which allows us to simplify the analysis greatly. In electromagnetics, this assumption permits us to decouple the Maxwell equations into two sets of fields or modes, and they are often called as: transverse magnetic and

transverse electric. Any field subject to the assumption of no variation in z can be written as the sum of these modes:

Transverse magnetic modes (TM_z), contain the following field components: $E_z(x, y, t)$, $H_x(x, y, t)$ and $H_y(x, y, t)$.

Transverse electric modes (TE_z), contain the following field components: $H_z(x, y, t)$, $E_x(x, y, t)$ and $E_y(x, y, t)$.

2D TM mode is [16]

$$\frac{\partial H_x}{\partial t} = \frac{1}{\mu}\left(-\frac{\partial E_z}{\partial y} - \rho' H_x\right) \tag{5.2a}$$

$$\frac{\partial H_y}{\partial t} = \frac{1}{\mu}\left(\frac{\partial E_z}{\partial x} - \rho' H_y\right) \tag{5.2b}$$

$$\frac{\partial E_z}{\partial t} = \frac{1}{\varepsilon}\left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} - \sigma E_z\right) \tag{5.2c}$$

2D TE mode is

$$\frac{\partial E_x}{\partial t} = \frac{1}{\varepsilon}\left(\frac{\partial H_z}{\partial y} - \sigma E_x\right) \tag{5.3a}$$

$$\frac{\partial E_y}{\partial t} = \frac{1}{\varepsilon}\left(\frac{\partial H_z}{\partial y} - \sigma E_y\right) \tag{5.3b}$$

$$\frac{\partial H_Z}{\partial t} = \frac{1}{\mu}\left(\frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial x} - \rho' H_z\right) \tag{5.3c}$$

where μ, ρ', ε, and σ are permeability, equivalent magnetic resistivity, permittivity, and conductivity, respectively.

TM and TE modes are decoupled, namely, they contain no common field vector components. In fact, these modes are completely independent for structures comprised of isotropic materials. That is, the modes can exist simultaneously with no mutual interactions. Problems having both TM and TE excitation can be solved by a superposition of these two separate problems [14].

When 2D TM mode is discretized, FDTD formulas are

$$H^{n+1/2}_{x,i,j+1/2} = D_a \cdot H^{n-1/2}_{x,i,j+1/2} + D_b\left(E^n_{z,i,j} - E^n_{z,i,j+1}\right) \tag{5.4a}$$

$$H^{n+1/2}_{y,i+1/2,j} = D_a \cdot H^{n-1/2}_{y,i+1/2,j} + D_b\left(E^n_{z,i+1,j} - E^n_{z,i,j}\right) \tag{5.4b}$$

$$E^{n+1}_{z,i,j} = C_a \cdot E^n_{z,i,j} + C_b\left(H^{n+1/2}_{y,i+1/2,j} - H^{n+1/2}_{y,i-1/2,j} + H^{n+1/2}_{x,i,j-1/2} - H^{n+1/2}_{x,i,j+1/2}\right) \tag{5.4c}$$

$$C_a = \frac{(2 \cdot \varepsilon - \sigma \cdot \Delta t)}{(2 \cdot \varepsilon + \sigma \cdot \Delta t)} \tag{5.5a}$$

$$C_b = \frac{(2 \cdot \Delta t)}{\Delta x \cdot (2 \cdot \varepsilon + \sigma \cdot \Delta t)} \tag{5.5b}$$

$$D_a = \frac{(2 \cdot \mu - \sigma^* \cdot \Delta t)}{(2 \cdot \mu + \sigma^* \cdot \Delta t)} \tag{5.5c}$$

$$D_b = \frac{(2 \cdot \Delta t)}{\Delta x (2 \cdot \mu + \sigma^* \cdot \Delta t)} \tag{5.5d}$$

where n denotes discrete time.

In a programming language, there is no location like $n + 1/2$. So these subscripts can be rounded to upper integer value [17], as followings.

$$H_{x,i,j+1}^{n+1} = D_a \cdot H_{x,i,j+1}^{n} + D_b\left(E_{z,i,j}^{n} - E_{z,i,j+1}^{n}\right) \tag{5.6a}$$

$$H_{y,i+1,j}^{n+1} = D_a \cdot H_{y,i+1,j}^{n-1/2} + D_b\left(E_{z,i+1,j}^{n} - E_{z,i,j}^{n}\right) \tag{5.6b}$$

$$E_{z,i,j}^{n+1} = C_a \cdot E_{z,i,j}^{n} + C_b\left(H_{y,i+1,j}^{n+1} - H_{y,i,j}^{n+1} + H_{x,i,j}^{n+1} - H_{x,i,j+1}^{n+1}\right) \tag{5.6c}$$

5.2.1 Perfectly Matched Layer ABC

Perfect matched layer (PML) ABC is an absorbing material boundary condition which is firstly proposed by J.P. Berenger. PML is proven very effective, reflectionless to all impinging waves (polarization, angles), and is also reflectionless over a broad-band.

According to Berenger PML technique, the computational area is surrounded by PML. The EM energy is absorbed rapidly in these layers so that perfect conductor can be set at the outmost. This can be also understood as that the interior area is matched to desired properties by the PML (Fig. 5.3).

For TM_z wave, E_z is split into E_{zx} and E_{zy}. And Faraday's Law and Ampere's Law break into four equations:

$$\varepsilon \frac{\partial E_{zx}}{\partial t} + \sigma_x E_{zx} = \frac{\partial H_y}{\partial x} \tag{5.7a}$$

$$\varepsilon \frac{\partial E_{zy}}{\partial t} + \sigma_y E_{zy} = -\frac{\partial H_x}{\partial y} \tag{5.7b}$$

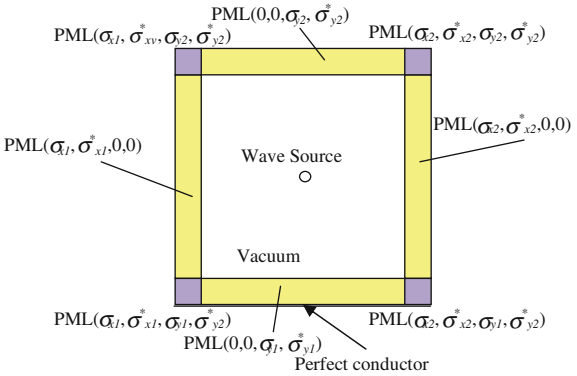


Fig. 5.3 The PML technique

$$\mu\frac{\partial H_x}{\partial t}+\sigma_y^* H_x=-\frac{\partial(E_{zx}+E_{zy})}{\partial y} \tag{5.7c}$$

$$\mu\frac{\partial H_y}{\partial t}+\sigma_x^* H_y=\frac{\partial(E_{zx}+E_{zy})}{\partial x} \tag{5.7d}$$

where σ^* is equivalent magnetic conductivity.

In PML area, the finite different equation is [18]:

$$\begin{aligned}H_x^{n+1}(i+1/2,j)=&\,e^{-\sigma_y^*(i+1/2,j)\delta t/\mu}H_x^n(i+1/2,j)\\&-\frac{(1-e^{-\sigma_y^*(i+1/2,j)\delta t/\mu})}{\sigma_y^*(i+1/2,j)\delta}\\&\times\begin{bmatrix}E_{zx}^{n+1/2}(i+1/2,j+1/2)+E_{zy}^{n+1/2}(i+1/2,j+1/2)\\-E_{zx}^{n+1/2}(i+1/2,j-1/2)-E_{zy}^{n+1/2}(i+1/2,j-1/2\end{bmatrix}\end{aligned} \tag{5.8a}$$

$$\begin{aligned}H_y^{n+1}(i,j+1/2)=&\,e^{-\sigma_x^*(i,j+1/2)\delta t/\mu}H_y^n(i,j+1/2)\\&-\frac{\left(1-e^{-\sigma_x^*(i.j+1/2)\delta t/\mu}\right)}{\sigma_x^*(i,j+1/2)\delta}\\&\times\begin{bmatrix}E_{zx}^{n+1/2}(i-1/2,j+1/2)+E_{zy}^{n+1/2}(i-1/2,j+1/2)\\-E_{zx}^{n+1/2}(i+1/2,j+1/2)-E_{zy}^{n+1/2}(i+1/2,j+1/2)\end{bmatrix}.\end{aligned} \tag{5.8b}$$

$$\begin{aligned}E_{zx}^{n+1/2}(i+1/2,j+1/2)=&\,e^{-\sigma_x(i+1/2,j+1/2)\delta t/\varepsilon}E_{zx}^{n-1/2}(i+1/2,j+1/2)\\&-\frac{(1-e^{-\sigma_x(i+1/2,j+1/2)\delta t/\varepsilon})}{\sigma_x(i+1/2,j+1/2)\delta}\\&\times\left[H_y^n(i,j+1/2)-H_y^n(i+1,j+1/2)\right]\end{aligned} \tag{5.8c}$$

$$E_{zy}^{n+1/2}(i+1/2,j+1/2) = e^{-\sigma_y(i+1/2,j+1/2)\delta t/\varepsilon} E_{zy}^{n-1/2}(i+1/2,j+1/2) - \frac{(1-e^{-\sigma_y(i+1/2,j+1/2)\delta t/\varepsilon}}{\sigma_y(i+1/2,j+1/2)\delta} \times \left[H_x^n(i+1/2,j+1) - H_x^n(i+1/2,j)\right] \quad (5.8d)$$

In PML, the magnetic and electric conductivity is matched so that there is not any reflection between layers. The wave impedance matching condition is [19]

$$\frac{\sigma}{\varepsilon} = \frac{\sigma^*}{\mu} \quad (5.9)$$

5.2.2 Mur's Absorbing Boundary Conditions

Spurious wave reflections occur at the boundaries of computational domain due to nature of FDTD code. Virtual absorbing boundaries must be used to prevent the reflections there. Many Absorbing boundary conditions (ABCs) have been developed over the past several decades. Mur's ABC is one of the most common ABCs. There are two types of Mur's ABC to estimate the fields on the boundary, which are first-order and second-order accurate. Mur's ABCs provide better absorption with fewer cells required between the object and the outer boundary, but at the expense of added complexity. The Mur's absorbing boundaries are adequate and relatively simple to apply [13].

FDTD simulations have been carried out in two dimensions with first order Mur's absorbing boundary conditions, therefore, they did not require a super computer system to perform. Considering the E_z component located at $x = i\Delta x$, $y = j\Delta y$ for 2D case, the first order Mur's estimation of E_z field component on the boundary is [17, 20]

$$E_{i,j}^{n+1} = E_{i-1,j}^{n} + \frac{c\Delta t - \Delta x}{c\Delta t + \Delta x}\left(E_{i-1,j}^{n+1} - E_{i,j}^{n}\right) \quad (5.10)$$

5.3 Developed Program

A program was developed in the Matlab programming language to examine the propagation of mobile phone radiation [21, 22]. Representation of the area analyzed in the program is shown in Fig. 5.4. Flow chart of the program is shown in Fig. 5.5. As shown in Fig. 5.5 firstly required input parameters of the program is entered by the user, and the area of analysis is divided into cells, and matrices are created for the electric and magnetic field components (E, H) calculated at each

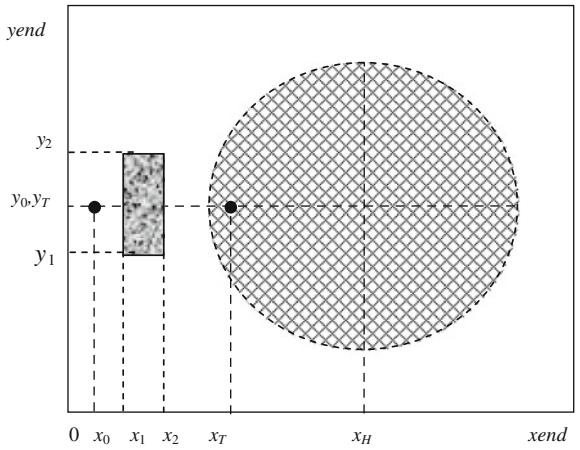


Fig. 5.4 Representation of the 2D simulation area

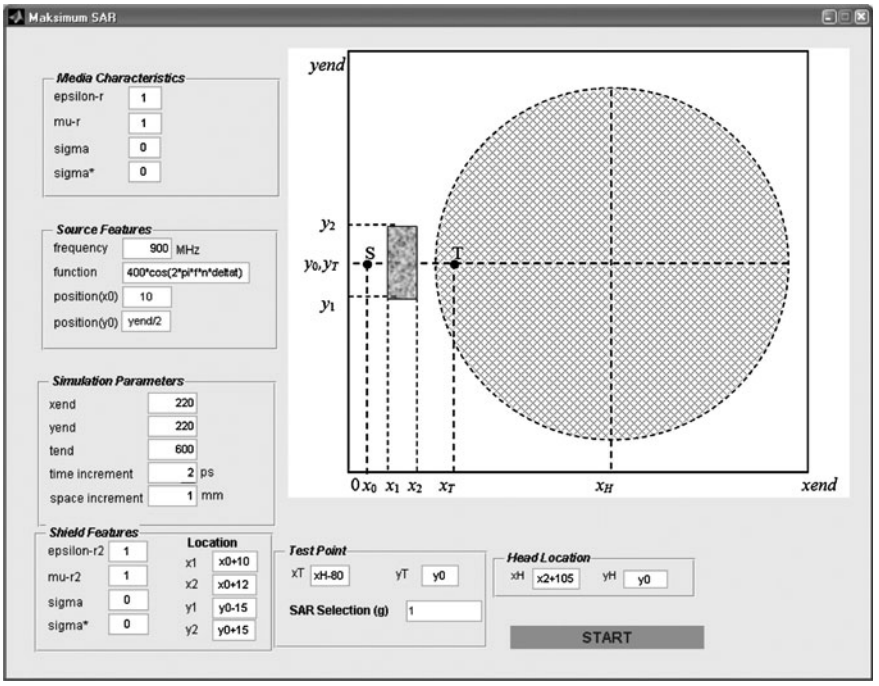


Fig. 5.5 Flow diagram for developed program

time step and each cell. Then, mathematical function of electric field emitted by mobile phone antenna is entered. Mur's absorbing boundary conditions are applied to eliminate artificial reflections and loops are carried out to calculate the electric and magnetic field values by stepping in the position and the time in the part that can be called FDTD Cycle.

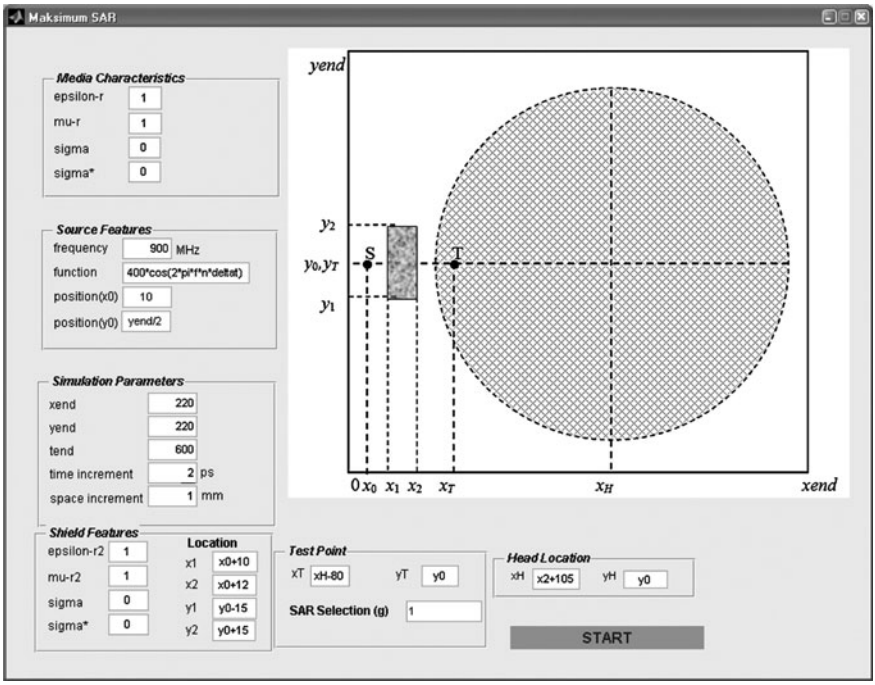


Fig. 5.6 Graphical interface of the developed program

The maximum electric field value is recorded at test point (T) for 1 or 10 g SAR. SAR values are calculated using the formula in Eq. 5.11 for each cell [23], and then 1 or 10 g averaged SAR is obtained by taking the average of them.

$$\mathrm{SAR} = \frac{\sigma |\mathrm{E_T}|^2}{2\rho} \quad (\mathrm{W/kg}) \tag{5.11}$$

Here, σ is average conductivity of the head, ρ is average mass density. E_T is the maximum electric field calculated for the test point.

A graphical interface has been designed for the developed program. This interface is shown in Fig. 5.6. All required data are entered here, and then the program is executed with the START button.

5.3.1 Input Parameters

Simulations were performed for unshielded case by entering the electrical properties of free space in Shield Features part in the developed program's graphical user interface and for shielded case by entering the electrical properties of copper ($\sigma = 5.8 \times 10^7$ S/m, 30×2 mm sized), separately. SAR was calculated at 8 cells

Table 5.1 Obtained SAR values from simulation results and shielding effectiveness values

	SAR (W/kg) without shield	SAR (W/kg) with copper shield	SE (dB)
1 g	0.7079	0.0061	−41.3
10 g	0.5958	0.0060	−39.9

for 1 g in the vicinity of test point, 80 cells for 10 g SAR. Output power of radiation source was assumed as constant during simulations. Average electrical conductivity of head in which SAR values were calculated was assumed as 0.97 S/m, the average mass density 1000 kg/m^3, relative permittivity 41.5, and the diameter 180 mm at 900 MHz [24, 25]. Time increment and space increment parameters of FDTD simulations were selected as 2 ps and 1 mm, respectively.

5.3.2 Simulation Results

1 g and 10 g averaged SAR values were calculated for both of cases as given in Table 5.1. Shielding effectiveness (SE) was calculated using the obtained values from simulation results with Eq. 5.12.

$$\mathrm{SE} = 20 \cdot \log \frac{s_1}{s_2} \quad (\mathrm{dB}) \tag{5.12}$$

Here, S_1 is the SAR value in shielded case, S_2 is one in unshielded case.

As shown in Table 5.1, SAR value decreased from 0.7079 W/kg to 0.0061 W/kg for 1 g averaged case and from 0.5958 to 0.0060 W/kg for 10 g averaged case under the effect of copper shield.

5.4 Conclusion

In this chapter, some information about mobile phones, their possible health risks, the parameter of SAR, its calculation and experimental measurement method, and numerical computing technique (2D-FDTD method) are expressed. In the application given in this chapter, reduction of radiation towards user from mobile phone with copper shield at 900 MHz frequency was investigated by calculating the SAR values in some simulations. The reason for choosing 2D and Mur's boundary condition in the simulation is to keep computer memory and processor requirements at minimum level. 1 and 10 g averaged SAR values were separately computed. In the simulations, shielding effectiveness was calculated using estimated SAR values for shielded and unshielded conditions.

As a result of simulations, it was found that the SAR values affecting mobile phone user were reduced about 40 dB by using copper shield.

Acknowledgments This work was supported by scientific research projects (BAP) coordinating office of Selçuk University.

References

1. Health Projection Agency [Online] Available: http://www.hpa.org.uk/Topics/Radiation/UnderstandingRadiation/UnderstandingRadiationTopics/ElectromagneticFields/MobilePhones/info_HealthAdvice/
2. Australian Radiation Protection and Nuclear Safety Agency [Online] Available: http://www.arpansa.gov.au/mobilephones/index.cfm
3. World Health Organization [Online] Available: http://www.who.int/mediacentre/factsheets/fs193/en/index.html
4. Ahlbom A, Bergqvist U, Bernhardt JH, Césarini JP, Court LA (1998) Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). Health Phys Soc 74(4):494–522
5. Kua L-C, Chuang H-R (2003) FDTD computation of fat layer effects on the SAR distribution in a multilayered superquadric-ellipsoidal head-model irradiated by a dipole antenna at 900/1800 MHz. In: IEEE International Symposium on Electromagnetic Compatibility
6. Kuo L-C, Lin C-C, Chunng H-R (2004) FDTD computation of fat layer effects on SAR distribution in a multilayered superquadric-ellipsoidal head model and MRI-based heads proximate to a dipole antenna. Radio Science Conference, Proceedings. Asia-Pacific, August 2004
7. Chen H-Y, Wang H-H (1994) Current and SAR induced in a human head model by the electromagnetic fields irradiated from a cellular phone. IEEE Trans Microwave Theory Tech 42(12):2249–2254
8. Schiavoni A, Bertotto P, Richiardi G, Bielli P (2000) SAR generated by commercial cellular phones-phone modeling, head modeling, and measurements. IEEE Trans Microwave Theory Tech 48(11):2064–2071
9. Kusuma AH, Sheta A-F, Elshafiey I, Alkanhal M, Aldosari S, Alshebeili SA (2010) Low SAR antenna design for modern wireless mobile terminals. In: STS International Conference, January 2010
10. Islam MT, Faruque MRI, Misran N (2009) Reduction of specific absorption rate (SAR) in the human head with ferrite material and metamaterial. Prog Electromagn Res C 9:47–58
11. Yee K (1966) Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans Antennas Propag 14:302–307
12. Taflove A, Brodwin ME (1975) Numerical solution of steady-state electromagnetic scattering problems using the time-dependent Maxwell's equations. IEEE Trans Microwave Theory Tech 23:623–630
13. Kunz KS, Luebbers RJ (1993) The finite difference time domain method for electromagnetism. CRC Press, Boca Raton
14. Stutzman WL, Thiele GA (1998) Antenna theory. Wiley, New York
15. Isaacson E, Keller HB (1967) Analysis of numerical methods. Wiley, New York
16. Davidson DB (2005) Computational electromagnetics for RF and microwave engineering. Cambridge University Press, Cambridge
17. Seyfi L, Yaldiz E (2006) Shielding analysis of mobile phone radiation with good conductors. In: Proceedings of the International Conference on Modeling and Simulation, vol 1, pp 189–194
18. Sadiku MNO (2000) Numerical techniques in electromagnetic, 2nd edn. CRC Press, Boca Raton
19. Berenger JP (1994) Matched layer for the absorption of electromagnetic waves. J Comp Phys 114:185–200 Aug
20. Mur G (1981) Absorbing boundary conditions for the finite-difference approximation of the time-domain electromagnetic field equations. IEEE Trans Electromag Compat 23:377–382
21. Seyfi L, Yaldız E (2010) Numerical computing of reduction of SAR values in a homogenous head model using copper shield. Lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, pp 839–843

22. Seyfi L, Yaldız E (2008) Simulation of reductions in radiation from cellular phones towards their users. In: First International Conference on Simulation Tools and Techniques for Communications. Networks and Systems, Marseille, France, March 2008
23. Foster KR, Chang K (eds) (2005) Encyclopedia of RF and microwave engineering, vol 1. Wiley-Interscience, Hoboken
24. Moustafa J, Abd-Alhameed RA, Vaul JA, Excell PS, McEwen NJ (2001) Investigations of reduced SAR personal communications handset using FDTD. In: Eleventh International Conference on Antennas and Propagation (IEE Conf Publ No 480), 17–20 April, pp 11–15
25. FCC, OET Bulletin 65c (2001) Evaluating compliance with FCC Guidelines for human exposure to radio frequency electromagnetic fields, [Online] Available: http://www.fcc.gov/Bureaus/Engineering_Technology/Documents/bulletins/oet65/oet65c.pdf

Chapter 6
Mitigation of Magnetic Field Under Overhead Transmission Line

Adel Zein El Dein Mohammed Moussa

Abstract The chapter presents an efficient way to mitigate the magnetic field resulting from the three-phase 500 kV single circuit overhead transmission line existing in Egypt, by using a passive loop conductor. The aim of this chapter is to reduce the amount of land required as right-of-way. The chapter used an accurate method for the evaluation of 50 Hz magnetic field produced by overhead transmission lines. This method is based on the matrix formalism of multiconductor transmission lines (MTL). This method obtained a correct evaluation of all the currents flowing in the MTL structure, including the currents in the subconductors of each phase bundle, the currents in the ground wires, the currents in the mitigation loop, and also the earth return currents. Furthermore, the analysis also incorporates the effect of the conductors sag between towers, and the effect of sag variation with the temperature on the calculated magnetic field. Good results have been obtained and passive loop conductor design parameters have been recommended for this system at ambient temperature (35°C).

6.1 Introduction

The rabid increase in HV transmission lines and irregular population areas near the manmade sources of electrical and magnetic fields, in Egypt, needs a suggestion of methods to minimize or eliminate the effect of magnetic and electrical fields on human begins in Egyptian environmental areas especially in irregular areas.

A. Z. El Dein Mohammed Moussa (✉)
Department of Electrical Engineering, High Institute of Energy, South Valley University, Aswan, 81258, Egypt
e-mail: azeinm2001@hotmail.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_6,
© Springer Science+Business Media B.V. 2011

Public concern about magnetic field effects on human safety has triggered a wealth of research efforts focused on the evaluation of magnetic fields produced by power lines [1–4]. Studies include the design of new compact transmission line configurations; the inclusion of auxiliary single or double lops for magnetic field mitigation in already existing power lines; the consideration of series-capacitor compensation schemes for enhancing magnetic field mitigation; the reconfiguration of lines to high phase operation, etc. [5–7]. However, many of the studies presented that deal with power lines make use of certain simplifying assumptions that, inevitably, give rise to inaccurate results in the computed magnetic fields. Ordinary simplifications include neglecting the earth currents, neglecting the ground wires, replacing bundle phase conductors with equivalent single conductors, and replacing actual sagged conductors with average height horizontal conductors. These assumptions result in a model where magnetic fields are distorted from those produced in reality [8, 9]. In this chapter, a matrix-based MTL model [10], where the effects of earth currents, ground wire currents and mitigation loop current are taken into account, is used; moreover, actual bundle conductors and conductors' sag at various temperatures are taken into consideration. The results from this method without mitigation loop are compared with those produced from the common practice method [8, 9] for magnetic field calculation where the power transmission lines are straight horizontal wires of infinite length, parallel to a flat ground and parallel with each other. Then the optimal design parameters of the mitigation loop for system under study are obtained.

6.2 Computation of System Currents

The MTL technique is used in this chapter for the simple purpose of deriving the relationship among the line currents of an overhead power line. This method is explained in [10], this chapter reviews and extends this method for Egyptian 500 kV overhead transmission line, with an other formula for the conductors' sag, taken into account the effect of temperature on the sag configuration [11]. The first step required to conduct a correct analysis consists in determination of all system currents based on prescribed phase-conductor currents I_p:

$$I_p = [I_1;\ I_2;\ I_3] \tag{6.1}$$

Consider the frequency-domain transmission line matrix equations for non-uniform MTLs (allowing the inclusion of the sag effect)

$$-\frac{dV}{dz} = Z'(\omega, z)I \tag{6.2a}$$

$$-\frac{dI}{dz} = Y'(\omega, z)V \tag{6.2b}$$

Where Z' and Y', denote the per-unit-length series-impedance and shunt-admittance matrices, respectively, V and I are complex column matrices collecting the phasors associated with all of the voltages and currents of the line conductors, respectively.

$$V = \begin{bmatrix} [V_a]_{1\times n_p} \\ [V_b]_{1\times n_p} \\ [V_c]_{1\times n_p} \\ [V_G]_{1\times n_G} \\ [V_L]_{1\times n_L} \end{bmatrix} \quad \text{and} \quad I = \begin{bmatrix} [I_a]_{1\times n_p} \\ [I_b]_{1\times n_p} \\ [I_c]_{1\times n_p} \\ [I_G]_{1\times n_G} \\ [I_L]_{1\times n_L} \end{bmatrix} \tag{6.3}$$

In (6.3), subscripts *a, b*, and *c* refer to the partition of phase bundles into three sub-conductor sets. Subscript G refers to ground wires and L subscript refers to the mitigation loop. In (6.3) n_p, n_G, and n_L denote, the number of phase bundles, the number of ground wires, and the number of conductors in the mitigation loop, respectively, for the Egyptian 500 kV overhead transmission line it is seen that: $n_p = 3$, $n_G = 2$, and $n_L = 2$ as it is proposed in this chapter. Since the separation of the electric and magnetic effects is an adequate approach for quasistationary regimes (50 Hz), where wave-propagation phenomena are negligible, all system currents are assumed to be Z independent. This means the transversal displacement currents among conductors are negligible or, in other words, (6.2b) equates to zero and only Z' values are needed to calculate. Since the standard procedure for computing Z' in (6.2a) has been established elsewhere [12–14], details will not be revealed here and thus only a brief summary is presented.

$$Z' = j\omega L + Z_E + Z_{skin} \tag{6.4}$$

The external-inductance matrix is a frequency-independent real symmetric matrix whose entries are:

$$L_{kk} = \frac{\mu_o}{2\pi} \ln \frac{2y_k}{r_k} \tag{6.5a}$$

$$L_{kk} = \frac{\mu_o}{4\pi} \ln \frac{(y_i + y_k)^2 + (x_i + x_k)^2}{(y_i - y_k)^2 + (x_i + x_k)^2} \tag{6.5b}$$

Where r_k denotes conductor radius, and y_k and x_k denote the vertical and horizontal coordinates of conductor k. Matrix Z_E, the earth impedance correction, is a frequency dependent complex matrix whose entries can be determine using Carson's theory or, alternatively, the Dubanton complex ground plane approach [12–14]. The entries of Z_E are defined as:

$$(Z_E)_{kk} = j\omega \frac{\mu_o}{2\pi} \ln\left(1 + \frac{\bar{P}}{y_k}\right) \tag{6.6a}$$

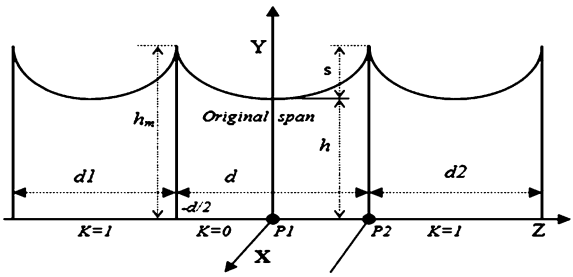


Fig. 6.1 Linear dimensions which determine parameters of the catenary

$$(Z_E)_{ik} = j\omega \frac{\mu_o}{4\pi} \ln\left(\frac{(y_i + y_k + 2\bar{P})^2 + (x_i - x_k)^2}{(y_i - y_k)^2 + (x_i - x_k)^2}\right) \tag{6.6b}$$

where $\bar{P}$, the complex depth, is given by $\bar{P} = (j\omega\mu_o/\rho)^{-1/2}$ with ρ denoting the earth resistivity. Matrix Z_{skin} is a frequency-dependent complex diagonal matrix whose entries can be determined by using the skin-effect theory results for cylindrical conductors [9]. For low-frequency situations, it will be:

$$(Z_{skin})_{kk} = (R_{dc})_k + j\omega \frac{\mu_o}{8\pi} \tag{6.7}$$

Where $(R_{dc})_k$ denotes the per-unit-length dc resistance of conductor k. Due to the line conductors' sag between towers; y_k will be a function on the distance z between the two towers, also the entries for L and Z_E, defined in (6.5a, b) and (6.6a, b), vary along the longitudinal coordinate z. The exact shape of a conductor suspended between two towers of equal height can be described by such parameters; as the distance between the points of suspension span, d, the sag of the conductor, S, the height of the lowest point above the ground, h, and the height of the highest point above the ground, hm. These parameters can be used in different combinations [13, 14]. Figure 6.1 depicts the basic catenary geometry for a single-conductor line, this geometry is described by:

$$y_k = h_k + 2\alpha_k \sinh^2\left(\frac{z}{2\alpha_k}\right) \tag{6.8}$$

Where α_k is the solution of the transcendental equation: $2[(hm_k - h_k)/d_k]u_k = \sinh^2(u_k)$, for conductor k; with $u_k = d_k/(4\alpha_k)$. The parameter α_k is also associated with the mechanical parameters of the line: $\alpha_k = (T_h)_k/w_k$ where $(T_h)_k$ is the conductor tension at mid-span and w_k is weight per unit length of the conductor k.

Consider a mitigation loop of length l, is present, where l is a multiple of the span length d. The line section under analysis has its near end at $-l/2$ and its far end at $l/2$. The integration of (6.2a) from $z = -l/2$ to $z = l/2$ gives:

$$V_{near} - V_{far} = I \int_{-l/2}^{l/2} Z'(z)dz \tag{6.9a}$$

Equation 6.9a can be written explicitly, in partitioned form, as:

$$\begin{bmatrix} \Delta V_a \\ \Delta V_b \\ \Delta V_c \\ \Delta V_G \\ \Delta V_L \end{bmatrix} = \begin{bmatrix} Z_{aa} & Z_{ab} & Z_{ac} & Z_{aG} & Z_{aL} \\ Z_{ba} & Z_{bb} & Z_{bc} & Z_{bG} & Z_{bL} \\ Z_{ca} & Z_{cb} & Z_{cc} & Z_{cG} & Z_{cL} \\ Z_{Ga} & Z_{Gb} & Z_{Gc} & Z_{GG} & Z_{GL} \\ Z_{La} & Z_{Lb} & Z_{Lc} & Z_{LG} & Z_{GL} \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \\ I_G \\ I_L \end{bmatrix} \tag{6.9b}$$

The computation of the bus impedance Z in Eq. 6.9a, b is performed using the following formula:

$$Z = \int_{-l/2}^{l/2} Z'(z)dz \tag{6.10}$$

Where values for Z' are evaluated from Eqs. 6.4–6.7 considering the conductors' heights given by (6.8). The two-conductor mitigation loop is closed and may include or not a series capacitor of impedance Z_c [7]. In any case, the submatrix I_L in (6.3) has the form:

$$I_L = \begin{bmatrix} I_{L1} \\ I_{L2} \end{bmatrix} = I_L S^T \tag{6.11}$$

where $S = [\,1 \quad -1\,]$.

By using the boundary conditions at both the near and far end of the line section, the voltage drop in the mitigation loop will be:

$$\Delta V_L = \begin{bmatrix} \Delta V_{L1} \\ \Delta V_{L2} \end{bmatrix} = \begin{bmatrix} V_{L1} \\ V_{L2} \end{bmatrix}_{near} - \begin{bmatrix} V_{L1} \\ V_{L2} \end{bmatrix}_{far} = \begin{bmatrix} \Delta V_{L1} \\ \Delta V_{L1} + Z_c I_L \end{bmatrix}$$

which can be written as:

$$S\Delta V_L = \Delta V_{L1} - \Delta V_{L2} = -Z_c I_L \tag{6.12}$$

Where I_L is the loop current, and $Z_c = jX_s = 1/(j\omega C_s)$ is the impedance of the series capacitor included in the loop.

Using (6.12), the fifth equation contained in (6.9b) allows for the evaluation of the currents flowing in the mitigation loop.

Using Eq. 6.12, the fifth equation contained in Eq. 6.9b allows for the evaluation of the currents flowing in the mitigation loop.

$$I_L = \begin{bmatrix} I_L \\ -I_L \end{bmatrix} = -\underbrace{YS^T SZ_{La}}_{K_{La}} I_a - \underbrace{YS^T SZ_{Lb}}_{K_{Lb}} I_b - \underbrace{YS^T SZ_{Lc}}_{K_{Lc}} I_c - \underbrace{YS^T SZ_{LG}}_{K_{LG}} I_G \quad (6.13)$$

Where; $Y = \frac{1}{(Z_c + SZ_{LL}S^T)}$.

Taking into account that the conductors belonging to given phase bundle are bonded to each other, and that ground wires are bonded to earth (tower resistances neglected), that result in: $\Delta V_a = \Delta V_b = \Delta V_c$ and $\Delta V_G = 0$.

By using $\Delta V_G = 0$ in the fourth equation contained in (6.9b) and using Eq. 6.13, the ground wire will be:

$$I_G = \underbrace{Y_G(Z_{Ga} - Z_{GL}K_{La})}_{K_{Ga}} I_a + \underbrace{Y_G(Z_{Gb} - Z_{GL}K_{Lb})}_{K_{Gb}} I_b + \underbrace{Y_G(Z_{Gc} - Z_{GL}K_{Lc})}_{K_{Gc}} I_c \quad (6.14)$$

Where; $Y_G = (Z_{GL}K_{LG} - Z_{GG})^{-1}$. Next, by using (6.13) and (6.14), I_L and I_G can be eliminated in (6.9b), yielding a reduced-order matrix problem

$$\begin{bmatrix} \Delta V_a \\ \Delta V_b \\ \Delta V_c \end{bmatrix} = \begin{bmatrix} \hat{Z}_{aa} & \hat{Z}_{ab} & \hat{Z}_{ab} \\ \hat{Z}_{ba} & \hat{Z}_{bb} & \hat{Z}_{bc} \\ \hat{Z}_{ca} & \hat{Z}_{cb} & \hat{Z}_{cc} \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \end{bmatrix} \quad (6.15)$$

Where;

$$\begin{aligned}
\hat{Z}_{aa} &= Z_{aa} + Z_{aG}K_{Ga} - Z_{aL}(K_{La} + K_{LG}K_{Ga}) \\
\hat{Z}_{ab} &= Z_{ab} + Z_{aG}K_{Gb} - Z_{aL}(K_{Lb} + K_{LG}K_{Gb}) \\
\hat{Z}_{ac} &= Z_{ac} + Z_{aG}K_{Gc} - Z_{aL}(K_{Lc} + K_{LG}K_{Gc}) \\
\hat{Z}_{ba} &= Z_{ba} + Z_{bG}K_{Ga} - Z_{bL}(K_{La} + K_{LG}K_{Ga}) \\
\hat{Z}_{bb} &= Z_{bb} + Z_{bG}K_{Gb} - Z_{bL}(K_{Lb} + K_{LG}K_{Gb}) \\
\hat{Z}_{bc} &= Z_{bc} + Z_{bG}K_{Gc} - Z_{bL}(K_{Lc} + K_{LG}K_{Gc}) \\
\hat{Z}_{ca} &= Z_{ca} + Z_{cG}K_{Ga} - Z_{cL}(K_{La} + K_{LG}K_{Ga}) \\
\hat{Z}_{cb} &= Z_{cb} + Z_{cG}K_{Gb} - Z_{cL}(K_{Lb} + K_{LG}K_{Gb}) \\
\hat{Z}_{cc} &= Z_{cc} + Z_{cG}K_{Gc} - Z_{cL}(K_{Lc} + K_{LG}K_{Gc})
\end{aligned}$$

The relationship between I_a, I_b and I_c is obtained from (6.15) by making $\Delta V_a = \Delta V_b = \Delta V_c$ and by using $I_a + I_b + I_c = I_p$. Then the following relations are obtained by:

$$I_a = KK_{ac}(KK_{ac} + K_{bc} + 1)^{-1} I_p \quad (6.16a)$$

$$I_b = KK_{bc}(KK_{ac} + K_{bc} + 1)^{-1} I_p \quad (6.16b)$$

$$I_c = (KK_{ac} + K_{bc} + 1)^{-1} I_p \quad (6.16c)$$

Table 6.1 Temperature effect

Temperature (°C)	15	20	25	30	35	40	45
Sag	7.3	7.8	8.3	8.8	9.3	9.8	10.3

Where; $KK_{ac} = KK_{ab}KK_{bc} + KK_{ac}$, $K_{ab} = Y_a\left(\hat{Z}_{bb} - \hat{Z}_{ab}\right)$, $K_{ac} = Y_a\left(\hat{Z}_{bc} - \hat{Z}_{ac}\right)$, $Y_a = \left(\hat{Z}_{aa} - \hat{Z}_{ba}\right)^{-1}$, $K_{bc} = (K_{bc1})^{-1}K_{cb1}$, $K_{bc1} = \hat{Z}_{ca}K_{ab} + \hat{Z}_{cb} - \hat{Z}_{ba}K_{ab} - \hat{Z}_{bb}$, and $K_{cb1} = \hat{Z}_{ba}K_{ac} + \hat{Z}_{bc} - \hat{Z}_{ca}K_{ac} - \hat{Z}_{cc}$.

Once I_P is given, all of the overhead conductor currents I_a, I_b, I_c, I_G and I_L can be evaluated, step after step using (6.13), (6.14), and (6.16a–c).

The net current returning through the earth I_E is the complement of the sum of all overhead conductor currents.

$$I_E = -\left[\sum_{k=1}^{n_p}(I_a)_k + \sum_{k=1}^{n_p}(I_b)_k + \sum_{k=1}^{n_p}(I_c)_k + \sum_{k=1}^{n_G}(I_G)_k + \sum_{k=1}^{n_L}(I_L)_k\right] \tag{6.17}$$

The sag of each conductor depends on individual characteristics of the line and environmental conditions. By using the Overhead Cable Sag Calculation Program [15], the variation of sag with temperatures can be calculated as in Table 6.1. Once all system currents are calculated, the magnetic field at any point, which produced from these currents, can be calculated.

6.3 Magnetic Field Calculations

By using the Integration Technique, which explained in details in [16] and reviewed here, the magnetic field produced by a multiphase conductors (M), and their images, in support structures at any point $P(x_o, y_o, z_o)$ can be obtained by using the Biot–Savart law as [9, 16]:

$$H_o = \frac{1}{4\pi}\sum_{K=1}^{M}\sum_{n=-N}^{N}\int_{-d/2}^{d/2}\left[(H_x)_k\vec{a}_x + (H_y)_k\vec{a}_y + (H_z)_k\vec{a}_z\right]dz \tag{6.18}$$

$$(H_x)_k = \frac{I_k\left[(z - z_o + nd)\sinh\left(\frac{z}{\alpha_k}\right) - (y_k - y_o)\right]}{d_k} - \frac{I_k\left[(z - z_o + nd)\sinh\left(\frac{z}{\alpha_k}\right) - (y_k + y_o + 2\bar{P})\right]}{d'_k} \tag{6.19}$$

$$(H_y)_k = \frac{I_k(x_k - x_o)}{d_k} - \frac{I_k(x_k - x_o)}{d'_k} \tag{6.20}$$

$$(H_z)_k = \frac{-I_k(x_k - x_o)\sinh\left(\frac{z}{\alpha_k}\right)}{d_k} + \frac{I_k(x_k - x_o)\sinh\left(\frac{z}{\alpha_k}\right)}{d'_k} \tag{6.21}$$

$$d_k = \left[(x_k - x_o)^2 + (y_k - y_o)^2 + (z - z_o + nd)^2\right]^{3/2} \tag{6.22}$$

$$d'_k = \left[(x_k - x_o)^2 + (y_k + y_o + 2\bar{P})^2 + (z - z_o + nd)^2\right]^{3/2} \tag{6.23}$$

The parameter (N) in (6.18) represents the number of spans to the right and to the left from the generic one where $K = 0$ as shown in Fig. 6.1.

6.4 Results and Discussion

The data used in the calculation of the magnetic field intensity at points one meter above ground level (field points), under Egyptian 500 kV TL single circuit are as presented in Table 6.2. The phase-conductor currents are defined by a balanced direct-sequence three-phase set of 50 Hz sinusoidal currents, with 2-kA rms, that is:

$$I_p = 2[1;\ \exp(-j2\pi/3);\ \exp(j2\pi/3)]\ \text{kA} \tag{6.24}$$

Figure 6.2 shows the effect of the number of spans (N) on the calculated magnetic field intensity. It is noticed that, when the magnetic field intensity calculated at point P1 (Fig. 6.1) and a distance away from the center phase, the effect of the spans' number is very small due to the symmetry of the spans around the calculation points, as explained in Fig. 6.1, where the contributions of the catenaries $d1$ and $d2$ are equal and smaller than the contribution of the catenary d, as

Table 6.2 Characteristics of 500 kV line conductors

Conductor number	Radius (mm)	X-Coordinate (m)	Y-Coordinate (m)	Rdc at 20°C (Ω/km)
1a	15.3	−13.425	22.13	0.0511
1b	15.3	−12.975	22.13	0.0511
1c	15.3	−13.2	21.74	0.0511
2a	15.3	−0.225	24.48	0.0511
2b	15.3	0.225	24.48	0.0511
2c	15.3	0	24.09	0.0511
3a	15.3	12.975	22.13	0.0511
3b	15.3	13.425	22.13	0.0511
3c	15.3	13.2	21.74	0.0511
G1	5.6	−8	30	0.564
G2	5.6	8	30	0.564
L1	11.2	−13.2	17	0.1168
L2	11.2	13.2	17	0.1168

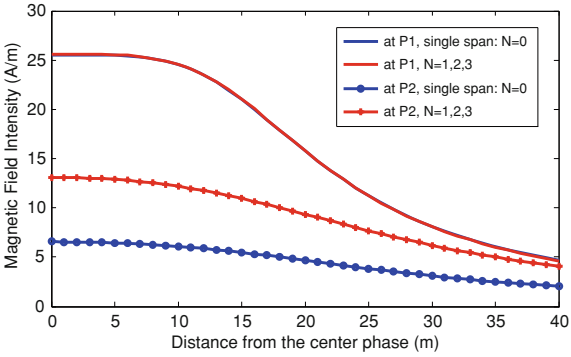


Fig. 6.2 The effect of the spans' numbers on the magnetic field intensity

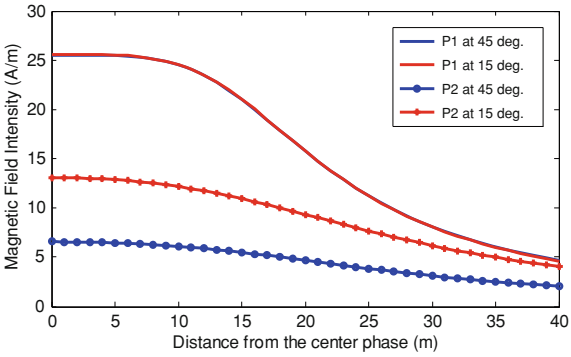


Fig. 6.3 The effect of the temperatures on the magnetic field intensity

they far from the field points. But when the magnetic field intensity calculated at point P2 (6.1) and a distance away from the center phase, the effect of the spans' number is of great effect (double), that due to the contribution of the catenary $d2$ which produced the same magnetic field intensity as the original span d in this case as explained in Fig. 6.1, and of course the catenary $d1$ have a small contribution in the calculated values of the magnetic field intensities in this case. Figure 6.3 shows the effects of the temperatures on the configuration of overhead transmission line conductors (sag) and hence on the calculated magnetic field intensity by using 3D integration technique with MTL technique. It is seen that as the sag increased with the increase in the temperatures (as indicated in Table 6.1), the magnetic field intensity also increased. Figure 6.4 shows the comparison between the magnetic filed calculated with both 2D straight line technique where the average conductors' heights are used, and 3D integration technique with MTL technique. It is seen that the observed maximum error of -23.2959% (at point P1) and 49.877% at (point P2) is mainly due to the negligence of the sag effect on the conductors.

Figure 6.5 shows the comparison between the magnetic field intensity calculated by using 3D integration technique with MTL technique with and without ground wires and with and without the short circuit mitigation loop. It is seen that, the observed maximum reduction of 1.9316% (at point P1) and 2.469%

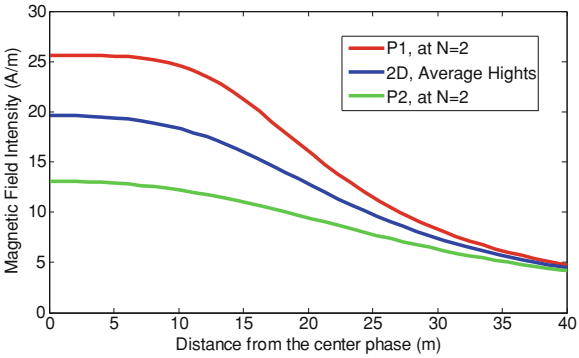


Fig. 6.4 Comparison between results of 2D and 3D techniques

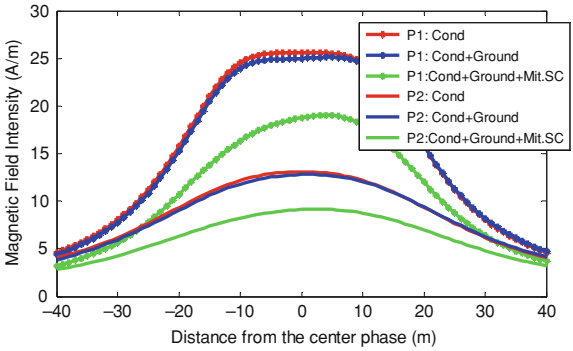


Fig. 6.5 Comparison between the calculated magnetic field from the conductors only; the conductors and ground wires; and the conductors, ground wires and S. C. mitigation loop

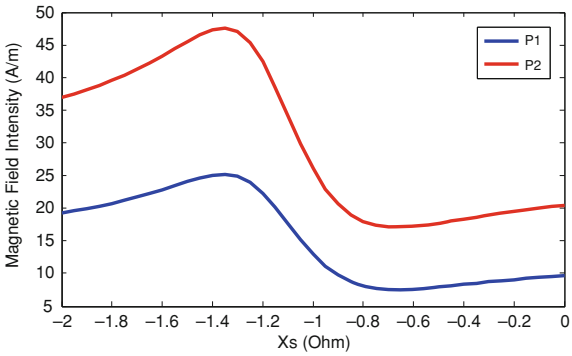


Fig. 6.6 Effect of the reactance Xs, inserted in the mitigation loop, on the calculated magnetic field

(at point P2) is mainly due to the negligence of the ground wires. It is seen that with the short circuit mitigation loop placed 5 m below beneath the outer phase conductors, the magnetic field intensity reduced to a significant values, maximum reduction of 25.7063% (at point P1) and 30.1525% (at point P2). The magnetic field intensity can be reduced further by inserting an appropriately chosen series

Table 6.3 The effect of the mitigation loop heights on the calculated magnetic field intensity at point (P1) and 26.4 m mitigation loop spacing

Height of mitigation loop	Magnetic field (A/m) at P1 at distance from center phase equals				
	−15 m	−10 m	0 m	10 m	15 m
18 m					
Short circuit	15.03	17.77	20.83	19.74	16.83
With optimal capacitance	9.42	10.80	17.52	18.84	16.1
19 m					
Short circuit	14.93	17.76	20.45	19.71	16.78
With optimal capacitance	8.88	10.82	17.12	18.77	15.98
20 m					
Short circuit	14.64	17.52	19.94	19.49	16.57
With optimal capacitance	8.13	10.43	16.63	18.64	15.84
21 m					
Short circuit	14.19	17.06	19.26	19.01	16.13
With optimal capacitance	7.01	9.56	15.87	18.15	15.40
23 m					
Short circuit	14.10	17.01	19.00	19.31	16.43
With optimal capacitance	7.07	9.86	16.35	19.16	16.41
24 m					
Short circuit	16.64	19.80	21.19	21.33	18.24
With optimal capacitance	11.46	14.13	17.97	19.79	16.96
25 m					
Short circuit	18.03	21.33	22.46	22.53	19.31
With optimal capacitance	13.95	16.87	19.55	20.85	17.91
26 m					
Short circuit	18.95	22.34	23.34	23.35	20.04
With optimal capacitance	15.70	18.764	20.82	21.80	18.74
27 m					
Short circuit	19.61	23.08	24.00	23.96	20.58
With optimal capacitance	16.98	20.16	21.84	22.58	19.42

capacitor in the mitigation loop, in order to determine the optimal capacitance Cs of the capacitor to be inserted in the mitigation loop, the magnetic field intensity calculated at point one meter above ground surface under center phase, considering different values of Zc where $Zc = jXs$, with the reactance Xs varies from -2Ω to 0.

Figure 6.6 shows the graphical results of the effect of the reactance Xs, inserted in the mitigation loop, on the magnetic field intensity, from which it is seen that the optimal situation (minimum value of magnetic field intensity) is characterized by $Cs = 4.897$ mF, and worst situation (maximum value of magnetic field intensity) is characterizes by $Cs = 2.358$ mF. (Tables 6.3 and 6.4) depict the effect of the mitigation loop height on the calculated magnetic field intensity at points P1 and P2, respectively, when the mitigation loop spacing is 26.4 m (exactly under the outer phases). It is seen that the optimal height is one meter below the outer phase

Table 6.4 The effect of the mitigation loop heights on the calculated magnetic field intensity at point (P2) and 26.4 m mitigation loop spacing

Height of Mitigation loop	Magnetic Field (A/m) at P2 at distance from center phase equals				
	−15 m	−10 m	0 m	10 m	15 m
18 m					
Short circuit	7.97	8.89	9.88	9.43	8.61
With optimal capacitance	5.49	6.28	7.77	7.99	7.44
19 m					
Short circuit	7.77	8.7	9.68	9.26	8.44
With optimal capacitance	5.15	5.98	7.53	7.85	7.30
20 m					
Short circuit	7.48	8.4	9.37	9.00	8.20
With optimal capacitance	4.67	5.54	7.21	7.64	7.12
21 m					
Short circuit	7.09	7.99	8.93	8.61	7.84
With optimal capacitance	3.98	4.88	6.67	7.25	6.76
23 m					
Short circuit	6.79	7.71	8.72	8.49	7.72
With optimal capacitance	3.91	4.93	6.98	7.72	7.22
24 m					
Short circuit	8.06	9.09	10.06	9.65	8.74
With optimal capacitance	5.52	6.49	8.01	8.30	7.66
25 m					
Short circuit	8.76	9.85	10.82	10.32	9.33
With optimal capacitance	6.67	7.68	8.99	9.01	8.25
26 m					
Short circuit	9.23	10.37	11.34	10.78	9.74
With optimal capacitance	7.5	8.56	9.76	9.61	8.75
27 m					
Short circuit	9.58	10.75	11.73	11.13	10.05
With optimal capacitance	8.13	9.22	10.37	10.09	9.17

conductors when the mitigation loop is short circuited and about one meter above the outer phase conductors when an optimal capacitance inserted in the mitigation loop. (Tables 6.5 and 6.6) depict the effect of the mitigation loop spacing on the calculated magnetic field intensity at points P1 and P2, respectively, when the mitigation loop height is 21 m. It is seen that the optimal spacing is the outer phase conductors spacing. Figure 6.7 shows the comparison between the calculated magnetic field intensity values result from; the conductors, ground wires and short circuit mitigation loop; and the conductors, ground wires and mitigation loop with optimal capacitance and optimal parameters obtained from (Tables 6.3, 6.4, 6.5 and 6.6). It is seen that the magnetic field intensity decreased further more, maximum reduction of 8.0552% (at point P1) and 19.5326% (at point P2).

Table 6.5 The effect of the mitigation loop spacings on the calculated magnetic field intensity at point (P1) and 21 m height

Distance of mitigation loop from the center phase	Magnetic field (A/m) at P1 at distance from center phase equals				
	−15 m	−10 m	0 m	10 m	15 m
5 m					
Short circuit	21.54	25.03	24.88	25.56	22.20
With optimal capacitance	20.77	24.07	23.28	24.83	21.75
7.5 m					
Short circuit	20.43	23.48	23.24	24.15	21.22
With optimal capacitance	18.65	21.21	20.67	22.73	20.26
10 m					
Short circuit	18.35	20.90	21.42	21.98	19.45
With optimal capacitance	14.77	16.58	18.25	20.19	18.01
13.2 m					
Short circuit	14.19	17.06	19.26	19.01	16.13
With optimal capacitance	7.01	9.56	15.87	18.15	15.40
15 m					
Short circuit	14.57	18.22	20.51	20.19	16.69
With optimal capacitance	7.66	11.28	17.14	19.17	16.12

Table 6.6 The effect of the mitigation loop spacings on the calculated magnetic field intensity at point (P2) and 21 m height

Distance of mitigation loop from the center phase	Magnetic field (A/m) at P2 at distance from center phase equals				
	−15 m	−10 m	0 m	10 m	15 m
5 m					
Short circuit	10.69	11.89	12.79	12.17	11.06
With optimal capacitance	10.21	11.32	12.16	11.73	10.72
7.5 m					
Short circuit	10.06	11.14	11.96	11.46	10.47
With optimal capacitance	9.01	9.94	10.73	10.57	9.77
10 m					
Short circuit	9.01	9.97	10.77	10.38	9.52
With optimal capacitance	7.13	7.93	8.91	9.05	8.44
13.2 m					
Short circuit	7.09	7.99	8.93	8.61	7.84
With optimal capacitance	3.98	4.88	6.67	7.25	6.76
15 m					
Short circuit	7.28	8.33	9.41	8.98	8.08
With optimal capacitance	4.34	5.38	7.24	7.69	7.10

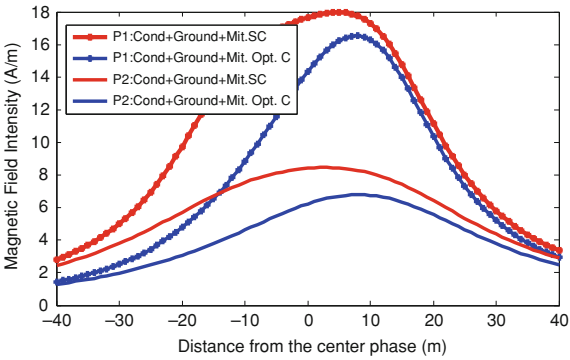


Fig. 6.7 Comparison between the calculated magnetic field intensity values result from the conductors, ground wires and short circuit mitigation loop; and from the conductors, ground wires and mitigation loop with capacitance of optimal value at optimal height and spacing

6.5 Conclusion

The effects of the currents in the subconductors of each phase bundle, the currents in the ground wires, the currents in the mitigation loop, and also the earth return currents; in the calculation of the magnetic field are investigated by using the MTL technique. Furthermore, the effect of the conductor's sag between towers, and the effect of sag variation with the temperature on the calculated magnetic field is studied. Finally the passive loop conductor design parameters, for Egyptian 500 kV overhead transmission line, are obtained at ambient temperature (35°C).

References

1. International Association of Engineers [Online]. Available: http://www.iaeng.org
2. El Dein AZ (2010) Mitigation of magnetic field under Egyptian 500 kV overhead transmission line. Lecture notes in engineering and computer science: Proceeding of the World Congress on Engineering, vol II WCE 2010, 30 June–2 July 2010, London, UK, pp 956–961
3. Hossam-Eldin AA (2001) Effect of electromagnetics fields from power lines on living organisms. In: IEEE 7th International Conference on Solid Dielectrics, June 25–29, Eindhoven, The Netherlands, pp 438–441
4. Karawia H, Youssef K, Hossam-Eldin AA (2008) Measurements and evaluation of adverse health effects of electromagnetic fields from low voltage equipments. MEPCON Aswan, Egypt, March 12–15, pp 436–440
5. Dahab AA, Amoura FK, Abu-Elhaija WS (2005) Comparison of magnetic-field distribution of noncompact and compact parallel transmission-line configurations. IEEE Trans Power Deliv 20(3):2114–2118
6. Stewart JR, Dale SJ, Klein KW (1993) Magnetic field reduction using high phase order lines. IEEE Trans Power Deliv 8(2):628–636
7. Yamazaki K, Kawamoto T, Fujinami H (2000) Requirements for power line magnetic field mitigation using a passive loop conductor. IEEE Trans Power Deliv 15(2):646–651
8. Olsen RG, Wong P (1992) Characteristics of low frequency electric and magnetic fields in the vicinity of electric power lines. IEEE Trans Power Deliv 7(4):2046–2053

9. Begamudre RD (2006) Extra high voltage AC. Transmission Engineering, third Edition, Book, Chapter 7, Wiley Eastern Limited, pp 172–205
10. Brandão Faria JA, Almeida ME (2007) Accurate calculation of magnetic-field intensity due to overhead power lines with or without mitigation loops with or without capacitor compensation. IEEE Trans Power Deliv 22(2):951–959
11. de Villiers W, Cloete JH, Wedepohl LM, Burger A (2008) Real-time sag monitoring system for high-voltage overhead transmission lines based on power-line carrier signal behavior. IEEE Trans Power Deliv 23(1):389–395
12. Noda T (2005) A double logarithmic approximation of Carson's ground-return impedance. IEEE Trans Power Deliv 21(1):472–479
13. Ramirez A, Uribe F (2007) A broad range algorithm for the evaluation of Carson's integral. IEEE Trans Power Deliv 22(2):1188–1193
14. Benato R, Caldon R (2007) Distribution line carrier: analysis procedure and applications to DG. IEEE Trans Power Deliv 22(1):575–583
15. Overhead Cable Sag Calculation Program http://infocom.cqu.edu.au/Staff/Michael_O_malley/web/overhead_cable_sag_calculator.html
16. El Dein AZ (2009) Magnetic field calculation under EHV transmission lines for more realistic cases. IEEE Trans Power Deliv 24(4):2214–2222

Chapter 7
Universal Approach of the Modified Nodal Analysis for Nonlinear Lumped Circuits in Transient Behavior

Lucian Mandache, Dumitru Topan and Ioana-Gabriela Sirbu

Abstract Recent approaches for time-domain analysis of lumped circuits deal with differential-algebraic-equation (DAE) systems instead of SPICE-type resistive models. Although simple and powerful, DAE models based on modified nodal approaches require some restrictions related to redundant variables or circuit topology. In this context, the paper proposes an improved version that allows treating nonlinear analog circuits of any topology, including floating capacitors, magnetically coupled inductors, excess elements and controlled sources. The procedure has been implemented in a dedicated program that builds the symbolic DAE model and solves it numerically.

7.1 Introduction

The transient analysis of analog nonlinear circuits requires a numerical integration that is commonly performed through associated discrete circuit models (SPICE-type models). In this manner, resistive circuits are solved sequentially at each time

L. Mandache (✉) · I.-G. Sirbu
Faculty of Electrical Engineering, University of Craiova, 107 Decebal Blv, 200440, Craiova, Romania
e-mail: lmandache@elth.ucv.ro

I.-G. Sirbu
e-mail: osirbu@elth.ucv.ro

D. Topan
Faculty of Electrical Engineering, University of Craiova, 13 A.I. Cuza Str, 200585, Craiova, Romania
e-mail: dtopan@central.ucv.ro

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_7,
© Springer Science+Business Media B.V. 2011

step [1–3]. Different strategies involve building of state or semistate mathematical models, as differential or differential–algebraic equation systems [4–6]. It is solved by specific numerical methods without engaging equivalent circuit models. Therefore, the problem of circuit analysis is transferred to a pure mathematical one. The latter strategy was extended during the last decades, taking advantage of the progress of the information technology [7–12].

The paper is focused on a semistate equations-based method associated to the modified nodal approach. This method avoids singular matrices in the equation system and overcomes the restriction related to floating capacitors [1, 14]. It also benefits by our previously developed topological analysis based on a single connection graph [3, 13] instead of two or more graphs [2, 10], although the circuit contains controlled sources. A simple, robust and comprehensive method is obtained.

The semistate mathematical model corresponding to the modified nodal approach (MNA) in the time domain has the general form of an ordinary differential equation system:

$$\begin{cases} M(x,t)\cdot\dot{x}(t)+N(x,t)\cdot x(t)=f(x,t)\,, \\ x(t_0)=x_0. \end{cases} \tag{7.1}$$

The vector of circuit variables $x(t)$, with the initial value $\mathbf{x}_0$, contains the vector of the node voltages $\mathbf{v}_{n-1}$ and the vector of the branch currents $\mathbf{i}_m$ that can not be expressed in terms or node voltages and/or theirs first-order derivatives:

$$x(t)=\begin{bmatrix} v_{n-1}(t) \\ i_m(t) \end{bmatrix}. \tag{7.2}$$

Therefore, the vector i_m contains the currents of zero-impedance elements (so-called MNA-incompatible elements): independent and controlled voltage sources, the controlling currents of current controlled sources, the inductor currents and the currents of the current controlled nonlinear resistors [1, 3, 5, 13].

$M(x,t)$ and $N(x,t)$ are square and generally state and time dependent matrices containing the parameters of the nonlinear elements. The matrix $\boldsymbol{M}$ contains the inductances and capacitances of energy storage circuit elements (dynamic parameters for the nonlinear storage elements) while the matrix $\boldsymbol{N}$ contains the resistances and conductances of resistors (dynamic parameters for the nonlinear resistors). Since the matrix $\boldsymbol{M}$ is commonly singular, the mathematical model (7.1) requires a special treatment.

$f(x,t)$ contains the circuit excitations and the parameters associated to the incremental sources used in the local linearization of the nonlinear resistors: current sources for voltage controlled nonlinear resistors and voltage sources for current controlled nonlinear resistors.

Although the building of the mathematical model (7.1) is relatively simple, the existence of a unique solution is debatable in most cases (the problem of possible singularity of the matrix $\boldsymbol{M}$ has been already reported).

The paper is organized as follows: the Sect. 7.2 explains the problem of floating capacitors related to the existence and uniqueness of the solution of the Eq. 7.1; the improved version of the MNA, in order to obtain an equivalent well-posed equation, is described in Sect. 7.3 and an example is treated in the Sect. 7.4.

7.2 The Problem of Floating Capacitors Related to MNA

The discussion on the floating capacitors requires building the circuit connection graph. We agree the single-graph procedure with its specific preliminary actions related to appropriate modeling of controlling ports of controlling sources [3]. Since the connection graph was built and the ground node was chosen, the capacitor subgraph is simply extracted.

If the circuit capacitors subgraph is not connected, then redundant variables appear in Eq. 7.1 and the circuit response can not be computed, as shown below.

The DAE mathematical model based on the MNA requires that any linear or nonlinear capacitor to be linked to the ground node through a path of capacitors. A capacitor that does not accomplish this requirement is called floating capacitor (see Fig. 7.1).

The time-domain nodal equations for such a structure are:

$$\begin{cases} C_k \cdot \dot{v}_p - C_k \cdot \dot{v}_q + \sum\limits_{j \in (p)} i_j = 0\,, \\ \\ -C_k \cdot \dot{v}_p + C_k \cdot \dot{v}_q + \sum\limits_{s \in (q)} i_s = 0\,, \end{cases} \tag{7.3}$$

where the state matrix

$$M_k = \begin{bmatrix} C_k & -C_k \\ -C_k & C_k \end{bmatrix}, \tag{7.4}$$

is obviously singular. Therefore, such a mathematical model is inappropriate. If one of the equations (7.3) is replaced by the cutset current law expressed for the cutset Σ surrounding the floating capacitor:

$$\begin{cases} C_k \cdot \dot{v}_p - C_k \cdot \dot{v}_q + \sum\limits_{j \in (p)} i_j = 0\,, \\ \\ \sum\limits_{\substack{j \in (p) \\ j \neq k}} i_j + \sum\limits_{\substack{s \in (q) \\ s \neq k}} i_s = 0\,, \end{cases} \tag{7.5}$$

then the singular matrix is avoided. The second equation in (7.5) can be obtained simply by adding both nodal equations (7.3). Nevertheless, the equation system (7.5) contains a redundant variable because two dynamic variables are involved by only one differential equation. Moreover, these two state variables correspond to

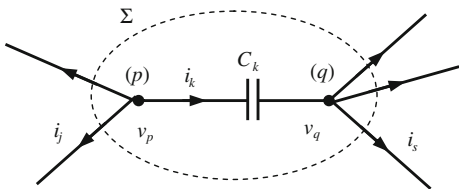


Fig. 7.1 Floating capacitor

only one capacitor. If the capacitor is grounded, with $v_q = 0$, then the nodal equation associated to the node (p) becomes

$$C_k \cdot \dot{v}_p + \sum_{j \in (p)} i_j = 0 \tag{7.6}$$

and the problem of singular matrix or redundant variable does not appear.

Extrapolating the above reasoning, if any subgraph of capacitors is floating, then the number of variables exceeds the number of essential capacitors, one variable being redundant.

In order to overcome the problem of redundant variable, a change of variables will be performed: the node voltages will be replaced by the tree-branch voltages.

7.3 Improved Version of the MNA

To overcome the problem of singular matrices and redundant variables introduced by the floating capacitors, our method requires accomplishing three main steps:

Step 1 Build the modified nodal equations by ignoring the floating capacitor problem;

Step 2 Identify all subgraphs of floating capacitors, as well as the nodal equations related to theirs nodes [1, 3]; for each such a subgraph, replace one of the nodal equations by the cutset current law expressed for the cutset surrounding the subgraph, as the second Eq. 7.5; an equivalent mathematical model is obtained, with the general form similar to (7.1):

$$M' \cdot \begin{bmatrix} \dot{v}_{n-1}(t) \\ \dot{i}_m(t) \end{bmatrix} + N' \cdot \begin{bmatrix} v_{n-1}(t) \\ i_m(t) \end{bmatrix} = f'. \tag{7.7}$$

Step 3 Perform a change of variables: the vector of the node voltages v_{n-1} is replaced by the vector of the tree-branch voltages u_t, the vector i_m remaining unchanged.

As it is known, the MNA does not require finding a normal tree of the given circuit. Nevertheless, in order to perform the change of variables a normal tree is required. We developed previously a simple and efficient method to build normal trees systematically [3, 13], that requires only few preliminary adjustments in the circuit diagram, as: the controlling branches of voltage-controlled sources must be modeled by zero-independent current sources and the controlling branches of current-controlled sources must be modeled by zero-independent voltage sources. The magnetically coupled inductors need to be modeled through equivalent diagrams with controlled sources. Thus, the normal tree is necessary for identifying the excess capacitors and inductors.

Since a normal tree was found, the step 3 of our algorithm can be performed. The node-branch incidence matrix is partitioned depending on the tree/cotree branches:

$$A = [A_t \quad | \quad A_c], \tag{7.8}$$

where A_t corresponds to the tree branches and A_c corresponds to the cotree branches [3, 13]. Next, the tree-branch voltages may be expressed in terms of nodes voltages [2, 13] using the transpose of the square nonsingular matrix A_t:

$$u_t = A_t^{\mathrm{t}} \cdot v_{n-1}. \tag{7.9}$$

Since the existence of the normal tree guarantees that the matrix A_t is always square and nonsingular, and consequently invertible, the node voltages of (7.9) can be expressed in terms of the tree-branch voltages:

$$v_{n-1} = A' \cdot u_t, \tag{7.10}$$

where A' signifies the inverse matrix of A_t^{t}. It is noticeable that the inverse matrix A' can be obtained relatively simple, due to the sparsity of A_t with the nonzero elements equal to $+1$ or -1.

Using (7.10) to substitute the vector v_{n-1} in (7.7), the mathematical model becomes

$$M' \cdot \begin{bmatrix} A' & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \dot{u}_t \\ \dot{i}_m \end{bmatrix} + N' \cdot \begin{bmatrix} A' & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} u_t \\ i_m \end{bmatrix} = f' \tag{7.11}$$

or

$$M''(x', t) \cdot \dot{x}'(t) + N''(x', t) \cdot x'(t) = f'(x', t) \tag{7.12}$$

where obvious notations were used. The new vector of variables is $x'(t)$. We extract from x' the essential capacitor voltages and the essential inductor currents, as elements of the state vector of length s (the subscript s comes from "state"):

$$x_s = \begin{bmatrix} u_C \\ i_L \end{bmatrix}. \tag{7.13}$$

The remained elements of x' are grouped in the vector x_a. The vector of variables organized as above involves splitting the equation system (7.12) as follows:

$$\left[\begin{array}{c|c} \boldsymbol{M}_{ss} & \boldsymbol{0}_{sa} \\ \hline \boldsymbol{0}_{as} & \boldsymbol{0}_{aa} \end{array}\right] \cdot \left[\begin{array}{c} \dot{\boldsymbol{x}}_s \\ \hline \dot{\boldsymbol{x}}_a \end{array}\right] + \left[\begin{array}{c|c} \boldsymbol{N}_{ss} & \boldsymbol{N}_{sa} \\ \hline \boldsymbol{N}_{as} & \boldsymbol{N}_{aa} \end{array}\right] \cdot \left[\begin{array}{c} \boldsymbol{x}_s \\ \hline \boldsymbol{x}_a \end{array}\right] = \left[\begin{array}{c} \boldsymbol{f}_s \\ \hline \boldsymbol{f}_a \end{array}\right]. \tag{7.14}$$

We remark that only the partition M_{ss} of size $s \times s$ of the matrix M'' is non-singular, all other elements being zeros.

A differential–algebraic equation system has been emphasized

$$\begin{cases} M_{ss} \cdot \dot{x}_s(t) + N_{ss} \cdot x_s(t) + N_{sa} \cdot x_a(t) = f_s(x_s, x_a, t)\,, \\ N_{as} \cdot x_s(t) + N_{aa} \cdot x_a(t) = f_a(x_s, x_a, t)\,, \end{cases} \tag{7.15}$$

with the initial condition

$$x_s(t_0) = \begin{bmatrix} u_C(t_0) \\ i_L(t_0) \end{bmatrix}. \tag{7.16}$$

Therefore, the vector x_s contains the variables of the differential equation system, while x_a contains the variables of the algebraic equation system (the subscript a comes from "algebraic").

In order to find the time-domain solution, many numerical techniques suitable for DAE can be used. In principle, the computation procedure requires the discretization of the analysis time and running the following steps:

- Solve the algebraic equation from (7.15), assigning to the state variables the initial values:

$$N_{as} \cdot x_s(t_0) + N_{aa} \cdot x_a = f_a \tag{7.17}$$

  in order to find the solution $x_a(t_0)$.
- Perform a numerical integration of the differential equation from (7.15), for the first discrete time interval $(t_0,\ t_1)$, assigning the value $x_a(t_0)$ to the vector x_a and considering (7.16) as initial condition:

$$\begin{cases} M_{ss} \cdot \dot{x}_s + N_{ss} \cdot x_s + N_{sa} \cdot x_a(t_0) = f_s\,, \\ x_s(t_0) = x_{s0}\,. \end{cases} \tag{7.18}$$

  The solution $x_s(t_1)$ is obtained.
- At the time step k the algebraic equation is solved, assigning to the state variables the values $x_s(t_k)$ calculated previously, during the numerical integration on the time interval $(t_{k-1},\ t_k)$:

$$N_{as} \cdot x_s(t_k) + N_{aa} \cdot x_a = f_a \,. \tag{7.19}$$

The solution $x_a(t_k)$ is found.

- Perform a numerical integration of the differential equation, for the next discrete time interval $(t_k,\ t_{k+1})$, assigning the previously computed value $x_a(t_k)$ to the vector x_a, and considering as initial condition the values $x_s(t_k)$:

$$\begin{cases} M_{ss} \cdot \dot{x}_s + N_{ss} \cdot x_s + N_{sa} \cdot x_a(t_k) = f_s \,, \\ x_s(t_k) = x_{sk} \,. \end{cases} \tag{7.20}$$

The solution $x_s(t_{k+1})$ is obtained.

The last two steps are repeated until the final moment of the analysis time is reached.

It is noticeable that the efficiency of the iterative algorithms used for nonlinear algebraic equation solving is significantly enhanced if $x_a(t_{k-1})$ is considered as start point.

The above described method has been implemented in a computation program under the high performance computing environment MATLAB. It recognizes the input data stored in a SPICE-compatible netlist, performs a topological analysis in order to build a normal tree and incidence matrices, identifies the excess elements and floating capacitors, builds the symbolic mathematical model as in expression (7.15), solves it numerically and represents the solution graphically.

7.4 Example

Let us study the transient behavior of an electromechanical system with a brushed permanent magnet DC motor supplied by a half wave uncontrolled rectifier, the mechanical load being nonlinear. The equivalent diagram built according to the transient model is shown in Fig. 7.2. There is not our goal to explain here the correspondence between the electromechanical system and the equivalent circuit diagram, or to judge the results from the point of view of its technical use. Only the algorithm described above will be emphasized.

The diagram contains one floating capacitor (branch 18) and two nonlinear resistors (the current controlled nonlinear resistor of the branch 11 is the model of the nonlinear mechanical load, reproducing the speed-torque curve, and the voltage controlled nonlinear resistor of the branch 16 corresponds to the semiconductor diode), whose characteristics are shown in Fig. 7.3. The independent zero-current sources 9, 10 and 14 correspond to the controlling branches of the voltage controlled sources 5, 13 and 7 respectively, while the independent zero-voltage source 4 corresponds to the controlling branch of the current controlled current source of the branch 6. The circuit is supplied by the independent sinusoidal voltage source of the branch 1.

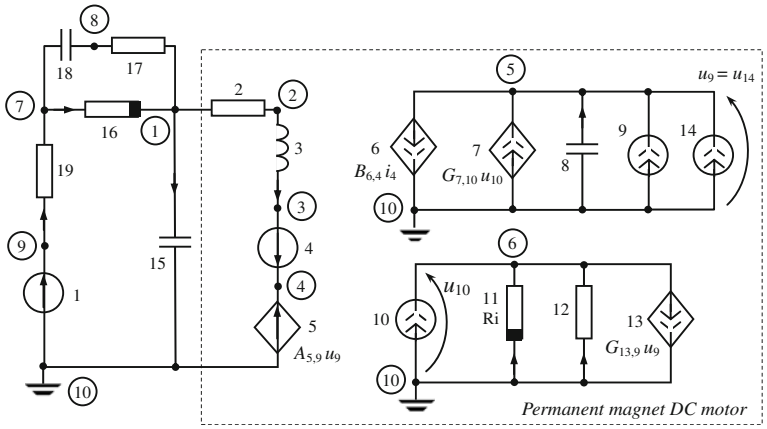


Fig. 7.2 Circuit example

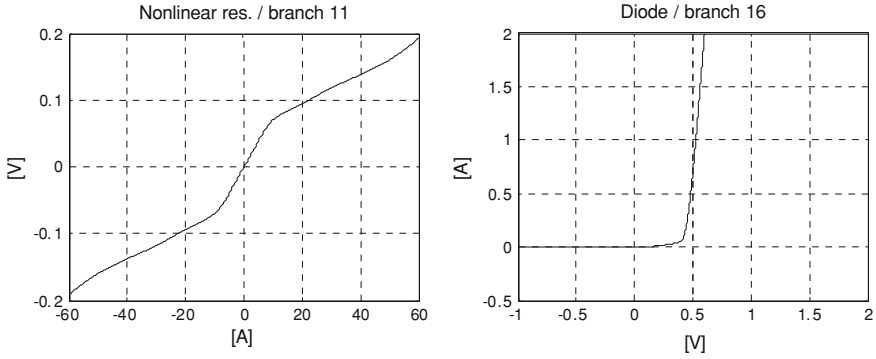


Fig. 7.3 Nonlinear resistors characteristics

If the node 10 is grounded, the topological analysis performed by our computing program gets the result:

The circuit does not contain excess inductors
Normal tree branches: 1 4 5 8 15 18 16 2 12
MNA-incompatible branches: 1 3 4 5 11
Floating capacitor subgraph 1:
reference_node: 8
other_nodes: 7

Therefore, the semistate variables are: $x_s = [u_8,\ u_{15},\ u_{18},\ i_3]^t$, and the variables of the algebraic equation system are: $x_a = [u_1,\ u_4,\ u_5,\ u_{16},\ u_2,\ u_{12},\ i_1,\ i_4,\ i_5,\ i_{11}]^t$.

The computing program gets the mathematical model in the symbolic form of type (7.15):

- The differential equation system:

$$\begin{cases} C15 * Du15 = -G17 * u18 - (-G17 - Gd16) * u16 - G2 * u2 + J0R16 \\ -C8 * Du8 = G7\_10 * u12 - B6\_4 * i4 + J9 + J14 \\ C18 * Du18 = -G19 * u15 - G19 * u1 - (Gd16 + G19) * u16 - J0R16 \\ -L3 * Di3 = -u15 + u4 - u5 + u2 \end{cases}$$

- The algebraic equation system:

$$\begin{cases} i3 - G2 * u2 = 0 \\ -i3 + i4 = 0 \\ -i4 - i5 = 0 \\ -G13\_14 * u8 + G12 * u12 + i11 + J10 = 0 \\ G19 * u15 - G17 * u18 + G19 * u1 + (G17 + Gd16 + G19) * u16 + J0R16 = 0 \\ -G19 * u15 - G19 * u1 - G19 * u16 - i1 = 0 \\ u1 + E \sin 1 = 0 \\ u4 + E4 = 0 \\ A5\_9 * u8 + u5 = 0 \\ u12 - Rd11 * i11 - E0R11 = 0 \end{cases}$$

Since the mathematical model above is given by the computing program automatically, some unobvious notations are used (e.g. the first derivative of the state variable u_{15} – $Du15$; the conductance of the nonlinear resistance of the branch 16 – $Gd16$; the incremental current source used in the local linearization of the nonlinear voltage-controlled nonlinear resistance of the branch 16 – $J0R16$; the voltage gain of the voltage-controlled voltage source of the branch 5 controlled by the branch 9 – $A5\_9$).

Assuming zero-initial conditions, the solving algorithm gets the result as time-domain functions, some examples being given in Fig. 7.4. Although the analysis time was 800 ms in order to cover the slowest component of the transient response, only details for the first 100 ms are shown in Fig. 7.4.

The DAE system has been solved using a Gear's numerical integration algorithm with variable time step combined with a Newton–Raphson algorithm. With the computation errors restricted to the limit values of 10^{-7} (absolute value) and 10^{-4} (relative value), the time step of the numerical integration process was maintained between 47 ns and 558 μs.

We remark that the same results have been obtained through a witness SPICE simulation, using the version ICAP/4 from Intusoft [15].

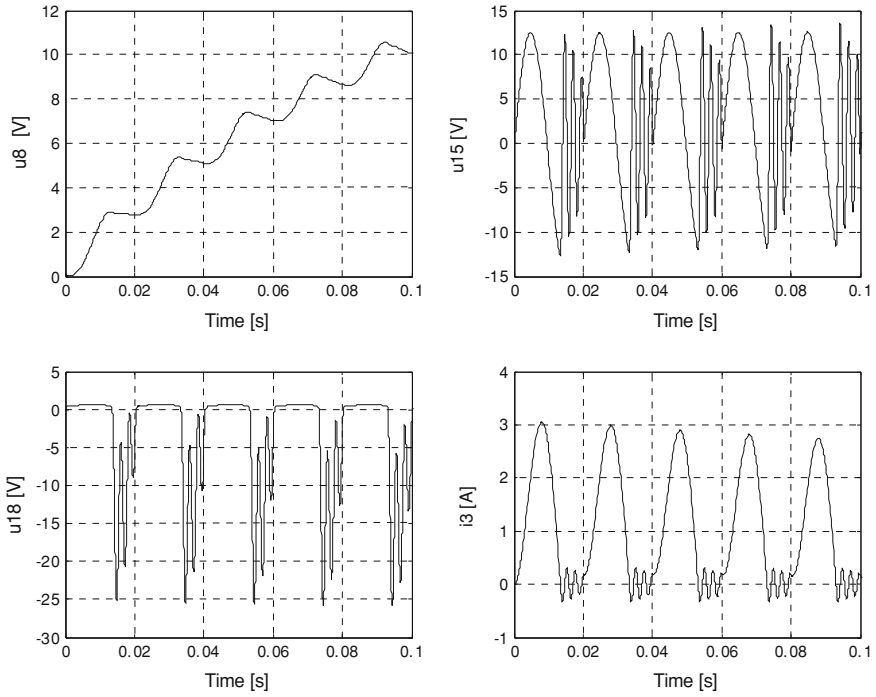


Fig. 7.4 Example of analysis results

7.5 Conclusion

An efficient and totally feasible algorithm intended to the time-domain analysis of nonlinear lumped analog circuits was developed and implemented in a computation program. It overcomes some restrictions of the modified nodal approaches, having practically an unlimited degree of generality for RLCM circuits.

The algorithm benefits by the simplicity of the MNA and the numerical methods for solving the mathematical model are flexible and can be optimized without requiring any companion diagrams (as the SPICE-like algorithms). In this manner, the computation time and the computer requirements can be reduced as compared to other methods. Our contribution is proven by an example, the chosen circuit containing nonlinear resistors, floating capacitors and controlled sources.

Acknowledgments This work was supported in part by the Romanian Ministry of Education, Research and Innovation under Grant PCE 539/2008.

References

1. Mandache L, Topan D (2010) Improved modified nodal analysis of nonlinear analog circuits in the time domain. Lecture notes in engineering and computer science, vol 2184—Proceedings of the World Congress on Engineering – London UK, June 30–July 2, pp 905–908
2. Chua LO, Lin PM (1975) Computer-aided analysis of electronic circuits–algorithms and computational techniques. Prentice-Hall, Englewood Cliffs
3. Iordache M, Mandache L (2004) Computer-aided analysis of nonlinear analog circuits (original title in Romanian) Ed. Politehnica Press, Bucharest (in Romanian)
4. Hodge A, Newcomb R (2002) Semistate theory and analog VLSI design. IEEE Circuits Syst Mag Second Quart 2(2):30–51
5. Newcomb R (1981) The semistate description of nonlinear time-variable circuits. IEEE Tram Circuits Syst CAS-28(1):62–71
6. Ho CW, Ruehli AE, Brennan PA (1975) The modified nodal approach to network analysis. IEEE Tram Circuits Syst CAS-22:504–509
7. Yamamura K, Sekiguchi T, Inoue Y (1999) A fixed-point homotopy method for solving modified nodal equations. IEEE Trans Circuits Syst - I: Fundam Theory Appl 46(6):654–665
8. Brambilla A, Premoli A, Storti-Gajani G (2005) Recasting modified nodal analysis to improve reliability in numerical circuit simulation. IEEE Trans Circuits Syst I: Regul Pap 52(3):522–534
9. Lee K, Park SB (1985) Reduced modified nodal approach to circuit analysis. IEEE Trans Circuits Syst 32(10):1056–1060
10. Chang FY (1997) The unified nodal approach to circuit analysis. In: IEEE International Symposium on Circuits and Systems, June 9–12, 1997, Hong Kong, pp 849–852
11. Hu JD, Yao H (1988) Generalized modified nodal formulation for circuits with nonlinear resistive multiports described by sample data. In: IEEE International Symposium on Circuits and Systems, vol 3, 7–9 June 1988, pp 2205–2208
12. Kang Y, Lacy JG (1992) Conversion of MNA equations to state variable form for nonlinear dynamical circuits. Electron Lett 28(13):1240–1241
13. Topan D, Mandache L (2007) Special matters of circuit analysis (original title in Romanian). Universitaria, Craiova (in Romanian)
14. Mandache L, Topan D (2003) An extension of the modified nodal analysis method. In: European Conference on Circuit Theory and Design ECCTD '03, September 1–4 2003, Kraków, pp II-410–II-413
15. *** ICAP/4—Is SPICE 4 User's guide (1998) Intusoft. San Pedro, California USA

Chapter 8
Modified 1.28 Tbit/s (32 × 4 × 10 Gbit/s) Absolute Polar Duty Cycle Division Multiplexing-WDM Transmission Over 320 km Standard Single Mode Fiber

Amin Malekmohammadi

Abstract A new version of Absolute Polar Duty Cycle Division Multiplexing transmission scheme over Wavelength Division Multiplexing system is proposed. We modeled and analyzed a method to improve the performance of AP-DCDM over WDM system by using Dual-Drive Mach–Zehnder-Modulator (DD-MZM). Almost 4.1 dB improvement in receiver sensitivity of 1.28 Tbit/s (32 × 40 Gbit/s) AP-DCDM-WDM over 320 km fiber is achieved by optimizing the bias voltage in DD-MZM.

8.1 Introduction

Wave length division multiplexing technologies have enabled the achievement of ultra high capacity transmission over 1 Tbit/s using Erbium Doped Fiber Amplifier (EDFA's). To pack a Tbit/s capacity into the gain bandwidth, spectral efficiency has to be improved. Narrow filtering characteristics and a high stability for the center frequency of optical filters are required to achieve dense WDM systems. Although such narrow optical filters could be developed [1, 2], narrow filtering of the signal light would result in wave form distortion in the received signal. Thus compact spectrum signals are also required for reducing distortion due to narrow filtering.

Absolute Polar Duty Cycle Division Multiplexing (AP-DCDM) is an alternative multiplexing technique which is able to support many users per WDM channel

A. Malekmohammadi (✉)
Department of Electrical and Electronic Engineering, The University of Nottingham, Malaysia Campus, Kuala Lumpur, Malaysia
e-mail: aminmalek_m@ieee.org

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_8,
© Springer Science+Business Media B.V. 2011

[3, 4]. Therefore, as reported in [4] the capacity of the WDM channels can be increased tremendously by using this technique. AP-DCDM enables us to use narrow optical filters that will provide spaces to increase the channel count. AP-DCDM system has intrinsic sensitivity penalty as compared to the binary signal, due to fragmentation of the main eye to smaller eyes [3]. At the same received power, these small eyes have different quality; therefore cause different AP-DCDM channels to have different performances, which is not desirable in telecommunication systems [3, 4].

In this paper, Dual-Drive Mach–Zehnder-Modulator (DD-MZM) is used in AP-DCDM-WDM setup at 1.28 Tbit/s transmission systems in order to improve the performance of AP-DCDM-WDM transmission system. It is shown that by optimum adjustment of the bias voltage at both ports, the sensitivity of the worst channel in AP-DCDM in 1.28 Tbit/s AP-DCDM-WDM over 320 km SSMF can be improved by 4.1 dB.

Mach–Zehnder modulators have the important feature that the chirp of the transmitted signal is a function of the electro optic properties of the p-i-n waveguide, the splitting ratios of the two branch waveguides, the differential length between the two arms of the interferometer, and the format of the modulating voltages applied to the arm electrodes [5–7]. An important property of DD–MZM is that, due to the quantum confined stark effect, the attenuation and phase constants of an optical signal propagating in the p-i-n waveguide are nonlinear functions of the applied voltage. Since these constants determine the modulator extinction ratio and chirp, the bias and modulation voltages can be optimized to yield the minimum degradation in receiver sensitivity due to fiber dispersion and self-phase modulation [6, 8].

8.2 Conventional 32-Channel AP-DCDM-WDM Transmission

As shown in Fig. 8.1 the evaluation starts with four AP-DCDM channels (4×10 Gbit/s) with PRBS of 2^{10}-1 (Fig. 8.1a) and followed by 32 WDM channels ($32 \times 4 \times 10$ Gbit/s) (Fig. 8.1b). In Fig. 8.1 four OOK channels were multiplexed by using AP-DCDM, whose outputs are multiplexed by using WDM technique (each WDM channel contains 4×10 Gbit/s with PRBS of 2^{10}-1 as shown in Fig. 8.1a).

62.5 GHz (0.5 nm) channel spacing was used. As a result, 128 AP-DCDM channels (32×4) are multiplexed in 32 WDM channels (λ_1 to λ_{32}) within $\sim$15.5 nm (1550–1565.5 nm) EDFA band. WDM spectral efficiency of 0.64 bit/s/Hz was achieved without polarization multiplexing [7]. The transmission line was 4 spans of 80 km Standard Single Mode Fiber (SSMF) followed by a 13.4 km Dispersion Compensation Fiber (DCF). The length ratio between SSMF and DCF is optimized so that the overall second-order dispersion reaches zero. For the SSMF, the simulated specifications for dispersion (D), dispersion slop (S), attenuation coefficient (α), effective area (A_{eff}) and nonlinear index of refraction (n_2) are

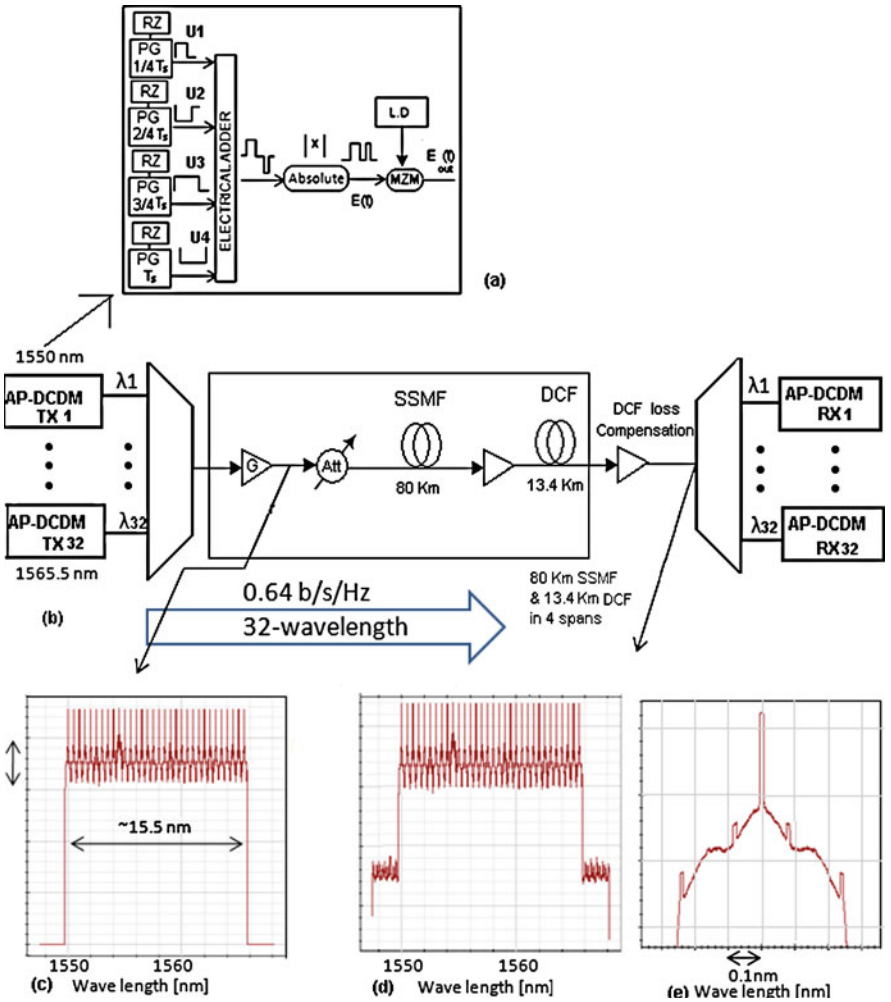


Fig. 8.1 **a** 4 × 10 Gbit/s AP-DCDM transmission system. **b** Simulation setup of 1.28 Tbit/s (32 × 4 × 10 Gbit/s) AP-DCDM-WDM transmissions. **c** Optical spectrum before transmission. **d** Optical spectrum after transmission. **e** Single channel AP-DCDM spectrum

16.75 ps/nm/km, 0.07 ps/nm^2/km, 0.2 dB/km, and 80 μm^2 and 2.7 × 10^{-20} m^2/W respectively. For DCF, D of ∼ −100 ps/nm/km, S of −0.3 ps/nm^2/km, α of 0.5 dB/km, A_{eff} of 12 μm^2 and n_2 of 2.6 × 10^{-20} m^2/W are used.

For Booster and pre-amplifier, an erbium-doped fiber amplifier (EDFA) with a flat gain of 30 dB and a noise figure (NF) of 5 dB was used. The total power to the booster is 8.35 dBm and launch power into SSMF is +15 dBm (∼0 dBm/channel). The Self Phase Modulation (SPM) effect in the link could be neglected since the launched power into the SSMF and DCF was less than the SPM threshold.

Figure 8.1c, d show the optical spectra of 32 WDM channels before and after transmission respectively. The effect of Four Wave Mixing (FWM) is negligible due to the phase mismatch in the highly dispersive transmission line [9, 10].

Figure 8.2a shows the exemplary eye diagrams taken after the 320 km SSMF (4 span of 80 km SSMF + 13.4 km DCF) for the worst channel (channel 16) of WDM system which contains 4 × 10 Gb/s AP-DCDM. As illustrated in Fig. 8.2 and reported in [7], the generated eye diagram for channel 16 which contains 4-channel of AP-DCDM system contains 6 small eyes. Eyes 1, 2, 3 and 4 (slots 1 and 2) correspond to the performance of AP-DCDM channel 1, eyes 2, 4 and 5 (slots 2 and 3) are related to performance of AP-DCDM channel 2, eyes 5 and 6 (slots 3 and 4) influence the performance of AP-DCDM channel 3, and eye 6 (slot 4) is related to AP-DCDM channel 4. As illustrated in Fig. 8.2a, at −25 dBm received power, Q-factor of all four eyes located at the first level is more than 6, which are higher than that of the eyes located at the second level (around 3.6 and 3.8 for eyes 1 and 2, respectively). The eye openings at different levels are almost similar but have different Q-factors due to different standard deviation of the noise variation at each level. Therefore, at the same received power, channel with minimum variation of noise has the best performance (e.g. channel 4) and the channel with maximum variation has the worst performance (channel 1).

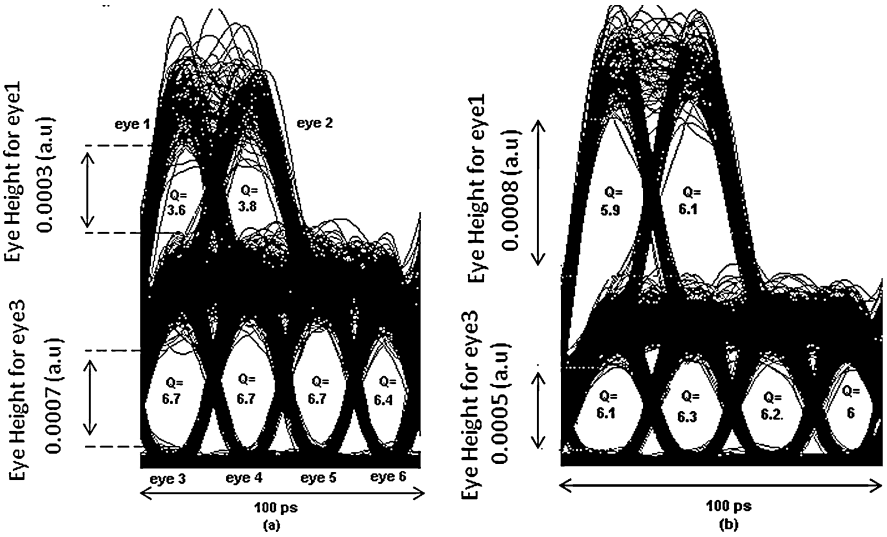


Fig. 8.2 **a** Received eye diagram for channel 16 in 32-channel AP-DCDM-WDM system. **b** Received eye diagram for channel 16 in 32-channel AP-DCDM-WDM system using optimized DD-MZM

8.3 Dual Drive-Mach–Zehnder Modulator

The Dual Drive-Mach–Zehnder modulator consists of an input Y-branch (splitter), two arms with independent drive electrodes, and an output Y-branch (combiner). The optical signal incident on the input Y-branch is split into the two arms of the interferometer. When the signals recombine at the output Y-branch, the on-state is achieved when there is no differential phase shift between the two signals and the off-state is achieved when there is a differential phase shift of radians.

The total optical field at the output of the Y-branch combiner is, to a good approximation, the sum of the fields at the outputs of the two arms. If the splitting ratios of the input and output Y-branches are identical, the output of the modulator is given by [6]

$$\begin{aligned} E(V_1, V_2) &= \frac{E_0}{1+SR}\left[SR.\exp\left(-\left(\frac{\Delta\alpha_\alpha(V_1)}{2}+j\Delta\beta(V_1)\right)L\right)\right] \\ &\quad + \exp\left(-\left(\frac{\Delta\alpha_\alpha(V_2)}{2}+j\Delta\beta(V_2)\right)L - j\Phi_0\right) \\ &= \sqrt{I(V_1,V_2)}\exp(j\varphi_0(V_1,V_2)) \end{aligned}$$

where $SR = P_1/P_2$ is a Y-branch power splitting ratio; $\Delta\alpha_\alpha/2$ is attenuation constant; $\Delta\beta$, phase constant; $\Phi_{0,}$ '0' radian for conventional modulator and 'Ω' radians for a Ω phase shift modulator; V_1 and V_2 are voltages applied to arms 1 and 2 respectively; I is the intensity of the optical signal; and Φ is the phase.

For $i = 1, 2$

$$V_i(t) = V_{bi} + V_{\text{mod}i}\, v(t)$$

where V_{bi} is the bias voltage; $V_{\text{mod}i}$ peak-to-peak modulation voltage; $V(t)$ modulation waveform with a peak-to-peak amplitude of one and an average value of zero.

The dependence of the attenuation and phase constants on the applied voltage can be obtained either by direct measurement of a straight section of waveguide cut from one arm of a modulator [5] or by using measurements of the voltage dependence of the intensity of the output signal for each arm with the other arm strongly absorbing [6–8].

Referring to Sect. 8.2, the improvement in the system performance can be obtained by having optimum amplitude distribution among the AP-DCDM signal level. This can be achieved by optimization in amplitude control of the level. To satisfy that requirement, we implement DD-MZM, which consists of an input Y-branch splitter, two arms with independent drive electrodes, and an output Y-branch combiner, in our setup as a replacement to conventional single-drive amplitude modulator (AM).

Table 8.1 DD-MZM optimization process for (a) Vb_2, (b) Vb_1

Setup	Q1	Q2	Q3	Q4	Q5	Q6
(a)						
Conventional AP-DCDM	3.6	3.8	6.7	6.7	6.7	6.4
MZM, $V_{b2} = -1$ v	4.4	4.7	6.3	6.5	6.4	6.1
MZM, $V_{b2} = -0.8$ v	5.1	5.4	6.2	6.4	6.2	6.1
MZM, $V_{b2} = -0.6$ v	5.9	6.1	6.1	6.3	6.2	6
MZM, $V_{b2} = -0.4$ v	6.4	5.9	6.9	6.1	6	5.8
(b)						
MZM, $V_{b1} = -3$ v	6.5	5.5	7	6.5	6	5.5
MZM, $V_{b1} = -2.9$ v	5.9	6.1	6.1	6.3	6.2	6
MZM, $V_{b1} = -2.8$ v	5.5	5.8	6.5	6.3	6.2	6
MZM, $V_{b1} = -2.6$ v	4.5	5	6.5	6.5	6.2	6

8.4 Optimizing the DD-MZM for 1.2 Tbit/s AP-DCDM-WDM Transmission

As discussed in Sect. 8.2 we need to have almost similar Q-factor for all 6 eyes to achieve similar performance for all channels. This can be done by improving the eye quality in second level. In order to change the eye high in second level while maintaining the maximum power, the bias voltage 1 (V_{b1}) and voltage 2 (V_{b2}) in DD-MZM need to be optimized so that the eye high in first level is reduced while increasing the eye high of the second level.

The optimum bias voltages are considered for two different conditions for the worst channel in 32 channel AP-DCDM-WDM system (Channel 16) as shown in Table 8.1. The dependence of Q-factor for all 6 eyes on the Vb_2 is shown in Table 8.1a (top) at the fixed received power of −25 dBm (receiver sensitivity of best channel), fixed $V_{b1}(-2.9)$ and splitting ratio of 1.3.

It can be seen from Table 8.1a that the optimum Vb2 is around −0.6 V where eye1 to eye 6 have almost similar Q-factors of 5.9, 6.1, 6.1, 6.3, 6.2, 6 respectively.

The variation of Q-factor for different values of the V_{b1} with fixed V_{b2} (−2.9) is shown in Table 8.1b. As illustrated in Table 8.1b, the optimum Vb_1 is around −2.9 where all eyes have similar Q-factor. Referring to Table 8.1 under optimized voltage biased conditions, the variation in the Q-factor is quite small and it is expected that the optimum sensitivity is essentially similar for all multiplexed channels [11].

8.5 32 Channels AP-DCDM-WDM System Performance Using Optimized DD-MZM

The simulation results are obtained by replacing AM in Fig. 8.1 by optimized DD-MZM for all 32 channels. The optimized DD-MZM was fixed with splitting

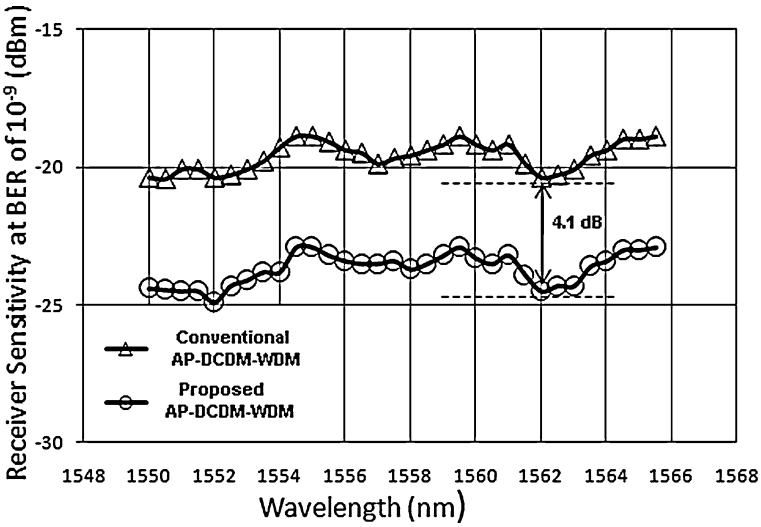


Fig. 8.3 Pre-amplified receiver sensitivity versus signal wavelength for 32 channels

ratio (SR) of 1.3, V_{b1} of −2.9 V and V_{b2} of −0.6 V. Figure 8.2b shows the exemplary eye diagrams taken after the 320 km SSMF (4 span of 80 km SSMF + 13.4 km DCF) for Channel 16.

As illustrated, although the eye highs are different, the Q-factors are almost the same. Compared to AP-DCDM with AM, Q-factors related to the second level are greatly improved (from 3.6 and 3.8 to 5.9 and 6.1 for eye 1 and 2 respectively). Note that the maximum amplitude values for AM and DD-MZM eye diagrams are the same. By improving the quality of the second level eyes, the performance of worse users (user 1 and 2) in middle WDM channel (Ch. 16) is significantly improved. In addition to that, we can have almost the same performance for all channels.

Figure 8.3 shows and compares the receiver sensitivity of both AP-DCDM-WDM with AM and the one with optimized DD-MZM for all 32 channels.

The degradation of receiver sensitivity is caused by the accumulated spontaneous emission light from each LD through the multiplexing process and by noise figure (NF) of the pre-amplifier. As shown in Fig. 8.3, the receiver sensitivity was around −21 dBm for conventional AP-DCDM-WDM system and the variation between the channels was around 1.5 dB. As illustrated in Fig. 8.3 the receiver sensitivity of proposed AP-DCDM-WDM system was improved to around −25.1 dBm compare to conventional AP-DCDM-WDM system. Therefore the proposed solution improves the receiver sensitivity by around 4.1 dB.

Figure 8.4 shows the improvement of OSNR for proposed AP-DCDM-WDM system compare to conventional AP-DCDM-WDM at BER of 10^{-9}. The reason

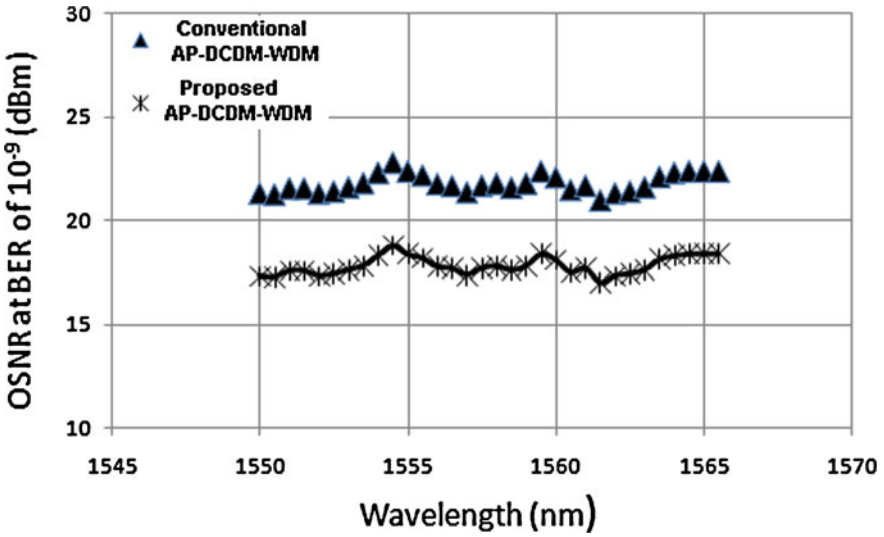


Fig. 8.4 OSNR versus signal wavelength for 32 channels

for this receiver sensitivity and OSNR improvement can be realized by looking and comparing the received eye diagrams depicted in Fig. 8.2a, b.

8.6 Conclusion

We have presented the performance of 1.28 Tbit/s AP-DCDM over WDM technique when drive voltages of DD-MZM are optimized. In comparison to the previous report [7], considerable receiver sensitivity improvement (4.1 dB) was achieved. The improvement is due to the eye high increment, which leads towards Q-factor enhancement. These results are impactful in the exploration for the optimum AP-DCDM transmission system.

References

1. Kim H, Essiambre R-J (2003) Transmission of 8 × 20 Gb/s DQPSK signals over 310-km SMF with 0.8-b/s/Hz spectral efficiency. IEEE Photon Technol Lett 15(5):769–771
2. Winzer P, Essiambre R (2006) Advance modulation formats for high-capacity optical transport networks. J Lightw Technol 24:4711–4728
3. Malekmohammadi A, Abdullah MK, Abas AF, Mahdiraji GA, Mokhtar M (2009) Analysis of RZ-OOK over absolute polar duty cycle division multiplexing in dispersive transmission medium. IET Optoelectron 3(4):197–206

4. Malekmohammadi A, Abas AF, Abdullah MK, Mahdiraji GA, Mokhtar M, (2009) Realization of high capacity transmission in fiber optic communication systems using absolute polar duty cycle division multiplexing (AP-DCDM) technique. Opt Fiber Technol 15(4):337–343
5. Cartledge C (1999) Optimizing the bias and modulation voltages of MQW Mach–Zehnder modulators for 10 Gb/s transmission on nondispersion shifted fiber. J Light Tech 17: 1142–1151
6. Adams DM, Rolland C, Fekecs A, McGhan D, Somani A, Bradshaw S, Poirier M, Dupont E, Cremer E, Anderson K (1998) 1.55 μm transmission at 2.5 Gbit/s over 1102 km of NDSF using discrete and monolithically integrated InGaAsP/InP Mach–Zehnder modulator and DFB laser. Electron Lett 34:771–773
7. Malekmohammadi A, Abas AF, Abdullah MK, Mahdiraji GA, Mokhtar M, Rasid M (2009) AP-DCDM over WDM system. Opt Commun 282:4233–4241
8. Hoshida T, Vassilieva O, Yamada K, Choudhary S, Pecqueur R, Kuwahara H (2002) Optimal 40 Gb/s modulation formats for spectrally efficient long-haul DWDM system. IEEE J Lightw Tech 20(12):1989–1996
9. Winzer PJ, Chandrasekhar S, Kim H (2003) Impact of filtering on RZ-DPSK reception. IEEE Photon Technol Lett 15(6):840–842
10. Suzuki S, Kawano Y, Nakasha Y (2005) A novel 50-Gbit/s NRZ-RZ converter with retiming function using Inp-HEMT technology. In: Presented at the Compound Semiconductor Integrated Circuit Symposium
11. Malekmohammadi A, Abdullah MK, Abas AF (2010) Performance, enhancement of AP-DCDM over WDM with dual drive Mach–Zehnder-Modulator in 1.28 Tbit/s optical fiber communication systems. Lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering 2010, WCE 2010, 30 June–2 July, London, UK, pp 948–951

Chapter 9
Wi-Fi Wep Point-to-Point Links

Performance Studies of IEEE 802.11a, b, g Laboratory Links

J. A. R. Pacheco de Carvalho, H. Veiga, N. Marques, C. F. Ribeiro Pacheco and A. D. Reis

Abstract The importance of wireless communications has been growing. Performance is a crucial issue, resulting in more reliable and efficient communications. Security is equally important. Laboratory measurements are made about several performance aspects of Wi-Fi (IEEE 802.11a, b, g) WEP point-to-point links. A contribution is given to performance evaluation of this technology, using two types of access points from Enterasys Networks (RBT-4102 and RBTR2). Detailed results are presented and discussed, namely at OSI levels 4 and 7, from TCP, UDP and FTP experiments: TCP throughput, jitter, percentage datagram loss and FTP transfer rate. Comparisons are made to corresponding results obtained for open links. Conclusions are drawn about the comparative performance of the links.

J. A. R. P. de Carvalho (✉) · C. F. Ribeiro Pacheco · A. D. Reis
Unidade de Detecção Remota, Universidade da Beira Interior, 6201-001 Covilhã, Portugal
e-mail: pacheco@ubi.pt

C. F. Ribeiro Pacheco
e-mail: a17597@ubi.pt

A. D. Reis
e-mail: adreis@ubi.pt

H. Veiga · N. Marques
Centro de Informática, Universidade da Beira Interior, 6201-001 Covilhã, Portugal
e-mail: hveiga@ubi.pt

N. Marques
e-mail: nmarques@ubi.pt

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_9,
© Springer Science+Business Media B.V. 2011

9.1 Introduction

Wireless communications are increasingly important for their versatility, mobility, speed and favourable prices. It is the case of microwave and laser based technologies, e.g. Wi-Fi (Wireless Fidelity) and FSO (Free Space Optics), respectively. The importance and utilization of Wi-Fi have been growing for complementing traditional wired networks. Wi-Fi has been used both in ad hoc mode, for communications in temporary situations e.g. meetings and conferences, and infrastructure mode. In this case, an AP (Access Point) is used to permit communications of Wi-Fi devices with a wired based LAN (Local Area Network) through a switch/router. In this way a WLAN, based on the AP, is formed which is known as a cell. A WPAN (Wireless Personal Area Network) arises in relation to a PAN (Personal Area Network).

Point-to-point and point-to-multipoint configurations are used both indoors and outdoors, requiring specialized directional and omnidirectional antennas. Wi-Fi uses microwaves in the 2.4 and 5 GHz frequency bands and IEEE 802.11a, 802.11b and 802.11g standards [1]. Due to increasing used of 2.4 GHz band, interferences increase. Then, the 5 GHz band has received considerable interest, although absorption increases and ranges are shorter.

Nominal transfer rates up to 11 (802.11b) and 54 Mbps (802.11a, g) are specified. CSMA/CA is the medium access control. Wireless communications, wave propagation [2, 3] and WLAN practical implementations [4] have been studied. Detailed information is available about the 802.11 architecture, including performance analysis of the effective transfer rate, where an optimum factor of 0.42 was presented for 11 Mbps point-to-point links [5]. Wi-Fi (802.11b) performance measurements are available for crowded indoor environments [6]. Performance has been a very important issue, giving more reliable and efficient communications. New telematic applications are specially sensitive to performances, when compared to traditional applications. Application characterization and requirements have been discussed e.g. for voice, Hi Fi audio, video on demand, moving images, HDTV images, virtual reality, interactive data, static images, intensive data, supercomputation, electronic mail, and file transfer [7]. E.g. requirements have been presented for video on demand/moving images (1–10 ms jitter and 1–10 Mbps throughputs) and for Hi Fi stereo audio (jitter less than 1 ms and 0.1–1 Mbps throughputs).

Wi-Fi microwave radio signals can be easily captured by everyone. WEP (Wired Equivalent Privacy) was initially intended to provide confidentiality comparable to that of a traditional wired network. In spite of its weaknesses, WEP is still widely used in Wi-Fi communications for security reasons. A shared key for data encryption is involved. In WEP, the communicating devices use the same key to encrypt and decrypt radio signals.

Several performance measurements have been made for 2.4 and 5 GHz Wi-Fi open links [8–10], as well as very high speed FSO [11]. In the present work further Wi-Fi (IEEE 802.11a, b, g) results arise, using WEP, through OSI levels 4 and 7.

Performance is evaluated in laboratory measurements of WEP point-to-point links using available equipments. Comparisons are made to corresponding results obtained for open links. Conclusions are drawn about the comparative performance of the links.

The rest of the paper is structured as follows: Chap. 2 presents the experimental details i.e. the measurement setup and procedure. Results and discussion are presented in Chap. 3. Conclusions are drawn in Chap. 4.

9.2 Experimental Details

Two types of experiments were carried out, which are referred as Exp-A and Exp-B. In the measurements of Exp-A we used Enterasys RoamAbout RBT-4102 level 2/3/4 access points (mentioned as AP-A), equipped with 16–20 dBm IEEE 802.11a/b/g transceivers and internal dual-band diversity antennas [12], and 100-Base-TX/10-Base-T Allied Telesis AT-8000S/16 level 2 switches [13]. The access points had transceivers based on the Atheros 5213A chipset, and firmware version 1.1.51. They were parameterized and monitored through both the console using CLI (Command Line Interface) and a HTTPS (Secure HTTP) incorporated server. The configuration was for minimum transmitted power and equivalent to point-to-point, LAN to LAN mode, using the internal antenna. For the measurements of Exp-B we used Enterasys RoamAbout RBTR2 level 2/3/4 access points (mentioned as AP-B), equipped with 15 dBm IEEE 802.11a/b/g cards [12], and 100-Base-TX/10-Base-T Allied Telesis AT-8000S/16 level 2 switches [13]. The access points had RBTBH-R2 W radio cards similar to the Agere-Systems model 0118 type, and firmware version 6.08.03. They were parameterized and monitored through both the console and the RoamAbout AP Manager software. The configuration was for minimum transmitted power i.e. micro cell, point-to-point, LAN to LAN mode, using the antenna which was built in the card.

Interference free channels were used in the communications. This was checked through a portable computer, equipped with a Wi-Fi 802.11a/b/g adapter running NetStumbler software [14]. WEP encryption was activated, using 128 bit encryption and a shared key for data encryption composed of 13 ASCII characters. No power levels above the minimum were required, as the access points were very close.

Both types of experiments, Exp-A and Exp-B, were made using a laboratory setup, which has been planned and implemented as shown in Fig. 9.1.

At OSI level 4 measurements were made for TCP connections and UDP communications, using Iperf software [15], permitting network performance results to be recorded. For a TCP connection, TCP throughput was obtained. For a UDP communication with a given bandwidth parameter, UDP throughput, jitter and percentage loss of datagrams were obtained. TCP packets and UDP datagrams of 1470 bytes size were used. A window size of 8 kbytes and a buffer size of the same value were used for TCP and UDP, respectively. In Fig. 9.1, one PC having

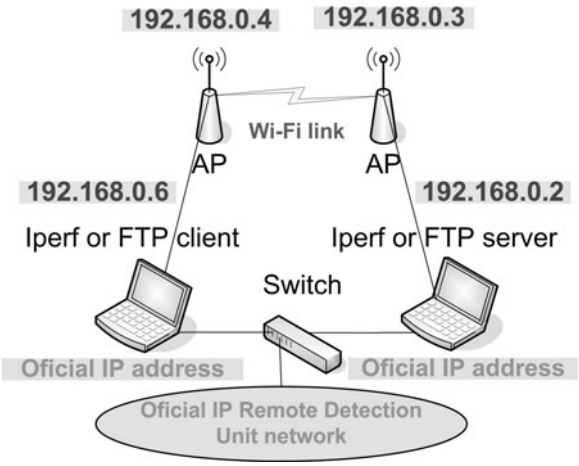


Fig. 9.1 Experimental laboratory setup scheme

IP 192.168.0.2 was the Iperf server and the other, with IP 192.168.0.6, was the Iperf client. Jitter, which represents the smooth mean of differences between consecutive transit times, was continuously computed by the server, as specified by RTP (Real Time Protocol) in RFC 1889 [16]. The same scheme was used for FTP measurements, where FTP server and client applications were installed in the PCs with IPs 192.168.0.2 and 192.168.0.6, respectively.

The PCs were portable computers running Windows XP. They were set up to make available maximum resources to the present work. Also, batch command files were written to enable the TCP, UDP and FTP tests. The results were obtained in batch mode and written as data files to the client PC disk. Each PC had a second network adapter, to permit remote control from the official IP Remote Detection Unit network, via switch.

9.3 Results and Discussion

Both access points AP-A and AP-B were configured with various fixed transfer rates for every one of the standards IEEE 802.11b (1, 2, 5.5 and 11 Mbps), 802.11g and 802.11a (6, 9, 12, 18, 24, 36, 48 and 54 Mbps).

At OSI level 1 (physical layer), for every one of the cases, the local and remote values of the signal to noise ratios SNR were recorded. The best SNR levels were observed for 802.11g and 802.11a.

Performance measurements, using TCP connections and UDP communications at OSI level 4 (transport layer), were carried out for both Exp-A and Exp-B. In each experiment, for every standard and nominal fixed transfer rate, an average TCP throughput was determined from several experiments. This value was used as the bandwidth parameter for every corresponding UDP test, giving average jitter and average percentage datagram loss. The results are shown in Figs. 9.2, 9.3 and 9.4.

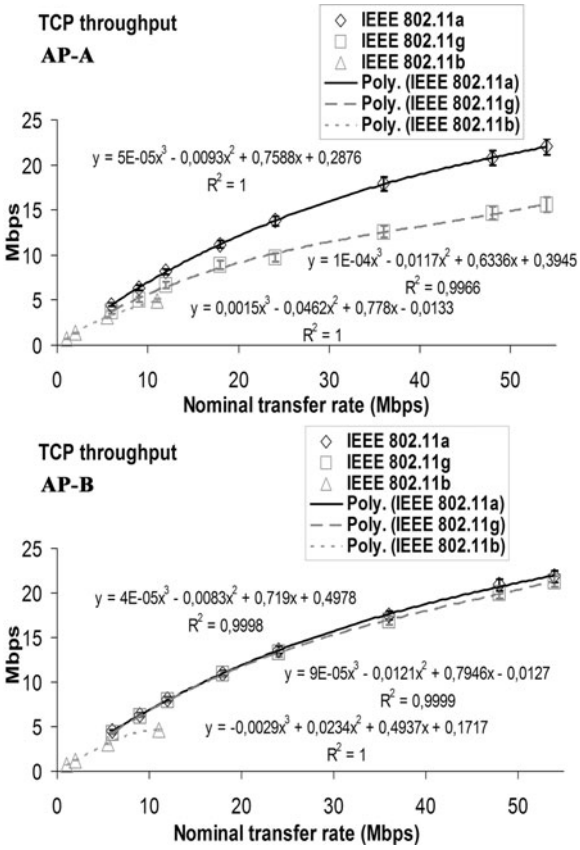


Fig. 9.2 TCP throughput results versus technology and nominal transfer rate; Exp-A and Exp-B

Figure 9.2 shows the results from Exp-A and Exp-B, where polynomial fits were made to the TCP throughput data for each AP implementation of IEEE 802.11a, b, g, where R^2 is the coefficient of determination. It follows that the best TCP throughputs are, by descending order, for 802.11a, 802.11g and 802.11b. In Exp-A (Fig. 9.2), the data for 802.11a are on average 32.6% higher than for 802.11g. The average values are 13.10 ± 0.39 Mbps for 802.11a, and 9.62 ± 0.29 Mbps for 802.11g. These values are in good agreement with those obtained for the same AP type and open links (13.19 ± 0.40 Mbps and 9.97 ± 0.30 Mbps for 802.11a and 802.11g, respectively) [9]. For 802.11b, the average value is 2.55 ± 0.08 Mbps. Also, the 802.11b data for 5.5 and 11 Mbps (average 4.05 ± 0.12 Mbps) are in good agreement with those obtained for the same AP type and open links (4.08 ± 0.12) [9]. In Exp-B (Fig. 9.2), the data for 802.11a are on average 2.9% higher than for IEEE 802.11g. The average values are 12.97 ± 0.39 Mbps for 802.11a, and 12.61 ± 0.38 Mbps for 802.11g. These values are in good agreement with those obtained for the same AP type and open links (12.92 ± 0.39 Mbps and 12.60 ± 0.38 Mbps for 802.11a and 802.11g, respectively) [9]. For 802.11b, the average value is 2.42 ± 0.07 Mbps.

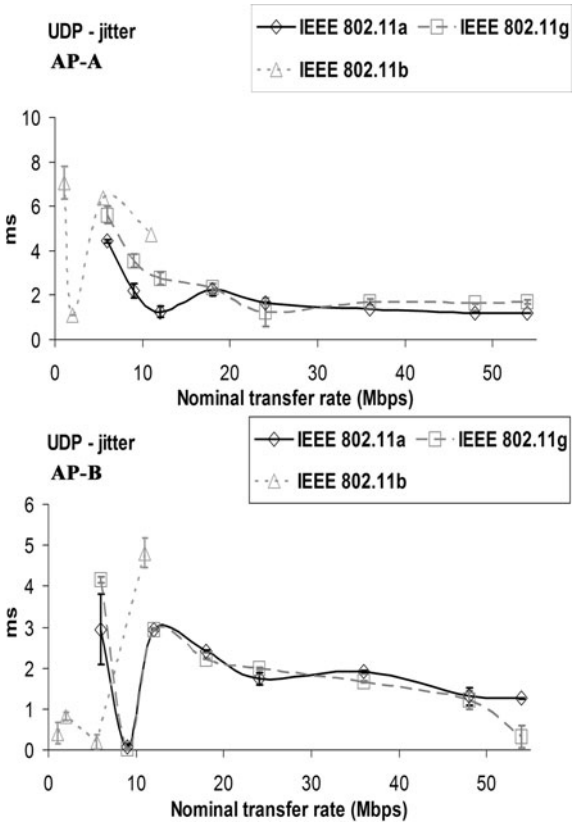


Fig. 9.3 UDP—jitter results versus technology and nominal transfer rate; Exp-A and Exp-B

Also, the 802.11b data for 5.5 and 11 Mbps (average 3.88 ± 0.12 Mbps) are in good agreement with those obtained for the same AP type and open links (3.84 ± 0.12) [9]. The best TCP throughput performance was for AP-B.

For both Exp-A and Exp-B, in Figs. 9.3 and 9.4 the data points representing jitter and percentage datagram loss, respectively, were joined by smoothed lines. In Exp-A (Fig. 9.3) the jitter data are on average lower for 802.11a (1.9 ± 0.1 ms) than for 802.11g (2.6 ± 0.1 ms). Similar trends were observed for the same AP type and open links (1.3 ± 0.1 ms for 802.11a and 2.4 ± 0.1 ms for 802.11g) [9]. For 802.11b, the average value is 4.8 ± 0.3 ms. Also, the 802.11b data for 5.5 and 11 Mbps (average 5.6 ± 0.9 ms) are higher than those respecting the same AP type and open links (3.7 ± 0.5 ms) [9]. In Exp-B (Fig. 9.3), the jitter data (1.8 ± 0.1 ms on average) show fair agreement for IEEE 802.11a and 802.11g. Similar trends were observed for the same AP type and open links (1.9 ± 0.1 ms on average) [9]. For 802.11b the average value is 1.6 ± 0.1 ms. Also, the 802.11b data for 5.5 and 11 Mbps (average 2.5 ± 0.5 ms) are in good agreement with those respecting the same AP type and open links (2.6 ± 0.2 ms) [9]. The best jitter performance was for AP-B.

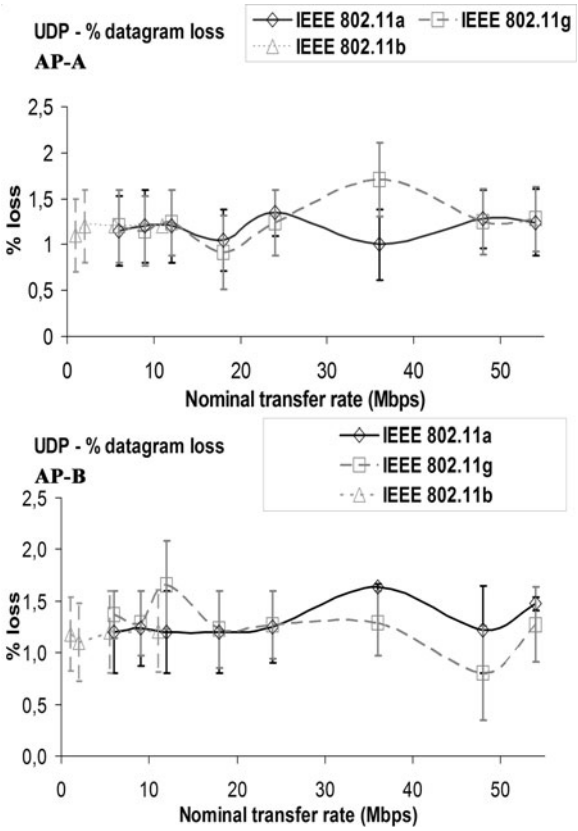


Fig. 9.4 UDP—percentage datagram loss results versus technology and nominal transfer rate; Exp-A and Exp-B

In both Exp-A and Exp-B (Fig. 9.4), generally, the percentage datagram loss data agree rather well for all standards. They are on average 1.2 ± 0.1%. This is in good agreement with the results for the same AP types and open links (on average, 1.3 ± 0.2% for AP-A and 1.2 ± 0.2% for AP-B) [9]. AP-A and AP-B have shown similar percentage datagram loss performances.

At OSI level 7 (application layer), FTP transfer rates were measured versus nominal transfer rates configured in the APs for the IEEE 802.11a, b, g, standards. Every measurement was the average for a single FTP transfer, using a binary file size of 100 Mbytes. The results from Exp-A and Exp-B are represented in Fig. 9.5. Polynomial fits to data were made for the implementation of every standard. It was found that in both cases the best performances were, by descending order, for 802.11a, 802.11g and 802.11b: the same trends found for TCP throughput. The FTP transfer rates obtained in Exp-A, using IEEE 802.11b, were close to those in Exp-B. The FTP performances obtained for Exp-A and IEEE 802.11a were only slightly better in comparison with Exp-B. On the contrary, for Exp-A and IEEE 802.11g, FTP performances were significantly worse than in Exp-B, suggesting that AP-B had a better FTP performance than AP-A for IEEE 802.11g. Similar trends had been observed for corresponding open links [9].

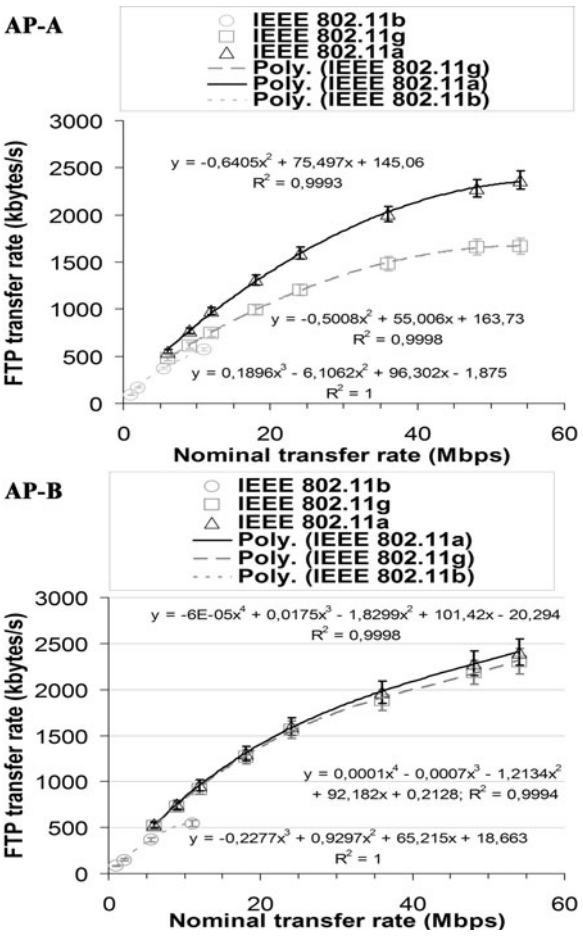


Fig. 9.5 FTP transfer rate results versus technology and nominal transfer rate; Exp-A and Exp-B

Generally, the results measured for the WEP links agree reasonably well, within the experimental errors, with corresponding data obtained for open links.

9.4 Conclusions

In the present work a laboratory setup arrangement was planned and implemented, that permitted systematic performance measurements of available access point equipments (RBT-4102 and RBTR2 from Enterasys) for Wi-Fi (IEEE 802.11a, b, g) in WEP point-to-point links.

Through OSI layer 4, TCP throughput, jitter and percentage datagram loss were measured and compared for each standard. The best TCP throughputs were found by descending order for 802.11a, 802.11g and 802.11b. TCP throughputs were also

found sensitive to AP type. Similar trends were observed for the same AP types and open links.

The lower average jitter values were found for IEEE 802.11a, and 802.11g. Some sensitivity to AP type was observed. For the percentage datagram loss, a reasonably good agreement was found, on average, for all standards and AP types. Similar trends were observed for the same AP types and open links.

At OSI layer 7, the measurements of FTP transfer rates have shown that the best FTP performances were by descending order for 802.11a, 802.11g and 802.11b. This result shows the same trends found for TCP throughput. Similar trends were observed for the same AP types and open links. FTP performances were also found sensitive to AP type.

Generally, the results measured for WEP links agree reasonably well, within the experimental errors, with corresponding data obtained for open links.

Additional performance measurements either started or are planned using several equipments, not only in laboratory but also in outdoor environments involving, mainly, medium range links.

Acknowledgments Supports from University of Beira Interior and FCT (Fundação para a Ciência e a Tecnologia)/POCI2010 (Programa Operacional Ciência e Inovação) are acknowledged. We acknowledge Enterasys Networks for their availability.

References

1. IEEE Std 802.11-2007, IEEE standard for local and metropolitan area networks-specific requirements-part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications (10 October 2007); http://standards.ieee.org/getieee802
2. Mark JW, Zhuang W (2003) Wireless communications and networking. Prentice-Hall Inc, Upper Saddle River
3. Rappaport TS (2002) Wireless communications principles and practice, 2nd edn. Prentice-Hall Inc, Upper Saddle River
4. Bruce WR III, Gilster R (2002) Wireless LANs end to end. Hungry Minds Inc, New York
5. Schwartz M (2005) Mobile wireless communications. Cambridge University Press, Cambridge
6. Sarkar NI, Sowerby KW (2006) High performance measurements in the crowded office environment: a case study. In: Proceedings ICCT'06-International Conference on Communication Technology, Guilin, China, 27–30 November, pp 1–4
7. Monteiro E, Boavida F (2002) Engineering of informatics networks, 4th edn. FCA-Editor of Informatics Ld, Lisbon
8. Pacheco de Carvalho JAR, Gomes PAJ, Veiga H, Reis AD (2008) Development of a university networking project. In: Putnik GD, Manuela Cunha M (eds) Encyclopedia of Networked and Virtual Organizations. IGI Global, Hershey, pp 409–422
9. Pacheco JAR, de Carvalho H, Veiga PAJ, Gomes CF, Ribeiro Pacheco N, Marques AD, Reis (2010) Wi-Fi Point-to-point Links—Performance Aspects of IEEE 802.11a, b, g Laboratory Links. In: Ao S-I, Gelman L (eds) Electronic Engineering and Computing Technology, Series: Lecture Notes in Electrical Engineering, vol 60. Springer, Netherlands, pp 507–514
10. Pacheco de Carvalho JAR, Veiga H, Marques N, Ribeiro Pacheco CF, Reis AD (2010) Laboratory performance of Wi-Fi WEP point-to-point links: a case study. Lecture notes in

engineering and computer science: Proceedings of the World Congress on Engineering 2010, WCE 2010, vol I. 30 June–2 July, London, UK, pp 764–767
11. Pacheco de Carvalho JAR, Veiga H, Gomes PAJ, Cláudia F, Ribeiro Pacheco FP, Reis AD (2008) Experimental performance study of very high speed free space optics link at the university of Beira interior campus: a case study. In: Proceedings ISSPIT 2008-8th IEEE International Symposium on Signal Processing and Information Technology, Sarajevo, Bosnia and Herzegovina, December 16–19, pp 154–157
12. Enterasys Networks, Roam About R2, RBT-4102 Wireless Access Points (20 December 2008). http://www.enterasys.com
13. Allied Telesis, AT-8000S/16 Layer 2 Managed Fast Ethernet Switch (20 December 2008). http://www.alliedtelesis.com
14. NetStumbler software. http://www.netstumbler.com
15. Iperf software, NLANR. http://dast.nlanr.net
16. Network Working Group, RFC 1889-RTP: A Transport Protocol for Real Time Applications. http://www.rfc-archive.org

Chapter 10
Interaction Between the Mobile Phone and Human Head of Various Sizes

Adel Zein El Dein Mohammed Moussa and Aladdein Amro

Abstract This chapter analyzes the specific absorption rate (SAR) induced in human head model of various sizes by a mobile phone at 900 and 1800 MHz. Specifically the study is considering in SAR between adults and children. Moreover, these differences are assessed for compliance with international safety guidelines. Also the effects of these head models on the most important terms for a mobile terminal antenna designer, namely: radiation efficiency, total efficiency and directivity, are investigated.

10.1 Introduction

In recent years, much attention has been paid to health implication of electromagnetic (EM) waves, especially human head part, which is exposed to the EM fields radiated from handsets. With the recent explosive increase of the use of mobile communication handsets, especially the number of children using a mobile phone, that develops many questions about the nature and degree of absorption of EM waves by this category of public as a function of their age and their morphology. For this reason the World Health Organization (WHO) has recommended

A. Z. E. D. M. Moussa (✉)
Department of Electrical Engineering, High Institute of Energy, South Valley University, Aswan, 81258, Egypt
e-mail: azeinm2001@hotmail.com

A. Amro
Department of Communications Engineering faculty of Engineering, Al-Hussein Bin Talal University, P.O. Box 20, Ma'an, Jordan
e-mail: amroru@hotmail.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_10,
© Springer Science+Business Media B.V. 2011

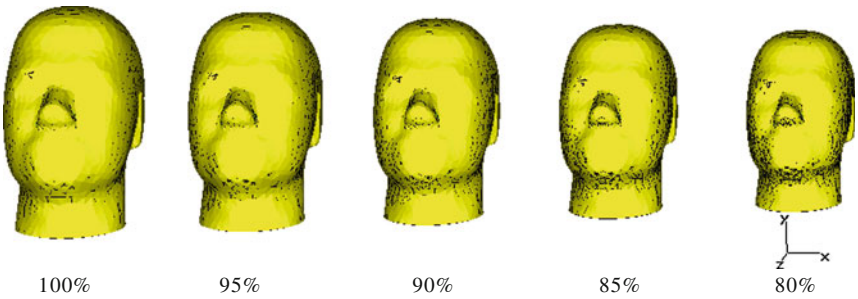


Fig. 10.1 Description of the various sizes of human head models

to undertake research studies on this subject [1–3]. This chapter investigates the effects of head models of various sizes on the most important terms for a mobile terminal antenna designer, namely: radiation efficiency, total efficiency and directivity; and also on the Specific Absorption Rates (SAR) which are induced in them. For this purpose, a comparison is performed concerning those parameters between an adult human head and some children heads obtained as a percent of an adult human head. The results are obtained using an electromagnetic field solver employing the Integral Equations method [4]. The SAR is the most appropriate metric for determining EM effect exposure in the very near field of a Radio Frequency (RF) source [5–9]. The local SAR (W/kg) at any point in the human head is defined as:

$$\mathrm{SAR} = \frac{\sigma E^2}{2\rho} \tag{10.1}$$

Where E is the peak amplitude of the electrical field in the human head tissue (V/m), σ is the tissue conductivity (S/m) and ρ is the tissue density (kg/m^3). The SAR over a mass of 10 and 1 g in the head and the other parameters of the mobile antenna are determined in each case.

10.2 Modeling of Human Head

For this study, five head models are used namely: that of an adult and other children human heads of sizes; 95, 90, 85, and 80% of the adult head size (which of size 100%), as they are shown in Fig. 10.1. Each head model consists of shell of skin tissue which is filled with a liquid of brain properties. For simulation of the EM fields in the human head, the appropriate parameters for the conductivity σ (S/m), the relative permittivity εr and the tissue density ρ (kg/m^3) of all different materials used for the calculation must be known. Additionally, the frequency dependence of these parameters must be considered and chosen appropriately. A recent compilation of Gabriel et al. covers a wide range of different body tissues

Table 10.1 Dielectric permittivity ε_r, conductivity σ (S/m), and mass density ρ (kg/m^3) of tissues used in the simulations at 900 and 1800 MHz

Properties of tissues		Dielectric permittivity ε_r	Conductivity σ (S/m)	Mass density ρ(kg/m^3)
Shell (skin)	900 MHz	43.8	0.86	1000
	1800 MHz	38.87	1.19	1000
Liquid (brain)	900 MHz	45.8	0.77	1030
	1800 MHz	43.5	1.15	1030

Table 10.2 Volume and mass of the heads' models

The volume and mass of the human head	Human head size as a percent of an adult one (%)				
	100	95	90	85	80
Tissue volume (mm^3)*106	5.5893	4.7886	4.0706	3.4283	2.8573
Tissue mass (kg)	5.7439	4.9236	4.1855	3.5250	2.9379

and offers equations to determine the appropriate dielectric values at each desired frequency [10, 11].

Table 10.1 shows the real part of the dielectric permittivity ε_r, conductivity σ (S/m), and mass density ρ (kg/m^3) of tissues used in the simulations at 900 and 1800 MHz. Table 10.2 shows the volume and the mass of the tissue of all children heads.

10.3 Modeling of the Mobile Phone

The mobile handset consists of a quarter-wavelength monopole (of radius 0.0025 m at 900 MHz and 0.001 m at 1800 MHz) mounted on a mobile handset (treated as a metal box of 1.8 × 4 × 10 cm), operates at 900 and 1800 MHz and radiated power of 0.125 W, as it is shown in Fig. 10.2.

10.4 Results and Discussion

Figures 10.3 and 10.4 present mobile terminal antenna designer parameters namely: return loss, radiation efficiency, total efficiency and directivity, the results obtained with the absence of the human head and at a frequency 1800 MHz. Table 10.3 present the Mobile Antenna Parameters, namely: radiation efficiency, total efficiency and directivity, the results obtained for various sizes of human heads and for the case of there absence. It is seen that as the size of human head decreases the radiation efficiency and total efficiency decrease, in the other side the directivity increases. The differences between the results of SAR of different kinds are given in Table 10.4 for each frequency and for each studied child head model.

Fig. 10.2 Description of the mobile handset

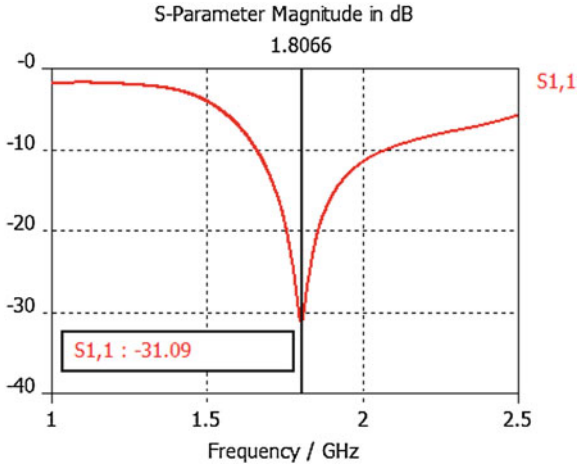


Fig. 10.3 Return loss without human head

The "SAR 10 grams" is the maximum SAR value averaged on 10 g which is obtained by averaging the SAR around each point in the volume and adding the nearest points till an average mass of 10 g is reached with a resulting volume having the shape of a portion of sphere. The "contiguous SAR 1 gram" is estimated by averaging the local maximum SAR, adding the highest SAR volume in a given tissue till a mass of 1 g is reached. The SAR (point) is the local value of SAR at every point inside the head model. The results show that by decreasing the head size the peak SAR 1 g and peak SAR 10 g decrease, however the percentage of absorbed power in the human head increases. So, the local SAR (point) and total SAR in children's heads increase as children's heads decrease, as indicated in Table 10.3. Also from Table 10.3 it is noticed that, the total SAR over the whole human head at 1800 MHz is less than that at 900 MHz. This is because the SAR regions produced by monopole antenna at 900 MHz are more extended as

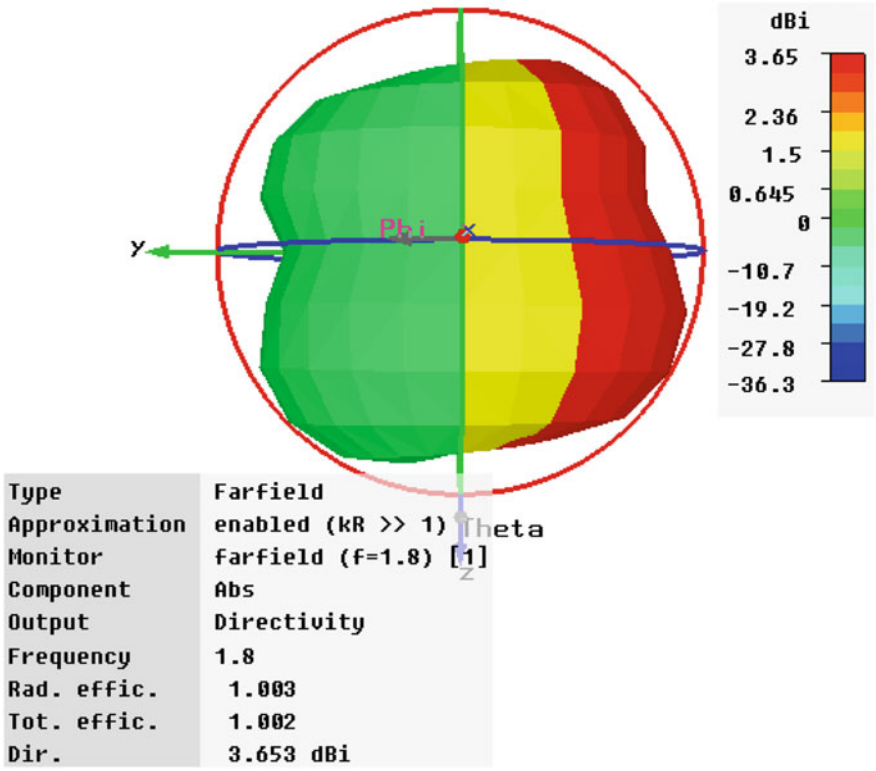


Fig. 10.4 Far field without human head

Table 10.3 Mobile antenna parameters with various sizes of human head

Mobile antenna parameters		Without human head	Human head size as a percent from an adult one (%)				
			100	95	90	85	80
900 MHz	Rad. η	1.003	0.276	0.292	0.311	0.335	0.357
	Tot. η	0.788	0.271	0.286	0.305	0.328	0.35
	Dir. (dBi)	2.627	6.066	5.943	5.859	5.819	5.712
1800 MHz	Rad. η	1.003	0.485	0.498	0.512	0.528	0.546
	Tot. η	1.002	0.476	0.489	0.504	0.521	0.539
	Dir. (dBi)	3.653	7.98	7.855	7.756	7.673	7.552

compared to those induced at 1800 MHz. The human body works as a barrier, mainly in high frequencies, because of skin depth. As the frequency increases the penetration capacity decreases and become more susceptible to obstacles.

Figures 10.5, 10.6, 10.7, 10.8, 10.9, 10.10 show the distributions of the local SAR, at the y = 0 plane; 10 g SAR in xz plane; and 1 g SAR in xy plane; in (W/kg), on the human head of various sizes, obtained with a radiated power of

Table 10.4 SAR induced in children's heads

Calculated parameters of the human head		Human head size as a percent from an adult one (%)				
		100	95	90	85	80
900 MHz	SAR (point)	1.134	1.206	1.124	1.122	1.214
	SAR 1 g	0.818	0.805	0.785	0.769	0.769
	SAR 10 g	0.593	0.59	0.584	0.58	0.572
	Absorbed power (wrms)	0.089	0.087	0.085	0.082	0.079
	Total SAR (W/kg)	0.016	0.018	0.02	0.023	0.027
1800 MHz	SAR (point)	4.149	3.078	2.404	2.319	2.282
	SAR 1 g	1.590	1.530	1.482	1.399	1.312
	SAR 10 g	0.922	0.887	0.848	0.805	0.764
	Absorbed power (wrms)	0.064	0.062	0.060	0.058	0.056
	Total SAR (W/kg)	0.011	0.012	0.014	0.016	0.019

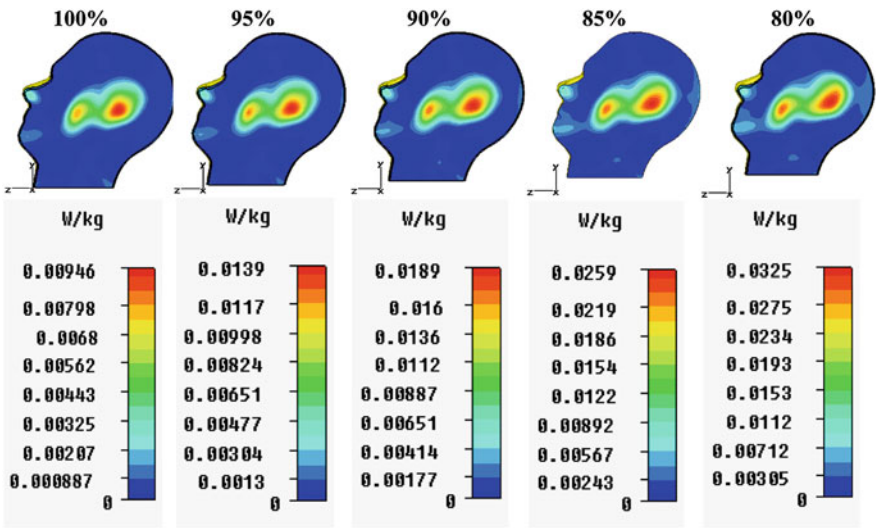


Fig. 10.5 Distributions of the local SAR at x = 0 plane for 1800 MHz

125 mW from a monopole antenna operates at 900 and 1800 MHz respectively. It can be easily noticed that high SAR regions produced by 900 MHz monopole antenna are more extended as compared to those induced by 1800 MHz monopole antenna, as it is explained before.

10.5 Conclusion

The obtained results show that the spatial-peak SAR values at a point or as averaged over 1 and 10 g on the human head of various sizes, obtained with a radiated power of 125 mW from a monopole antenna operates at 900 and

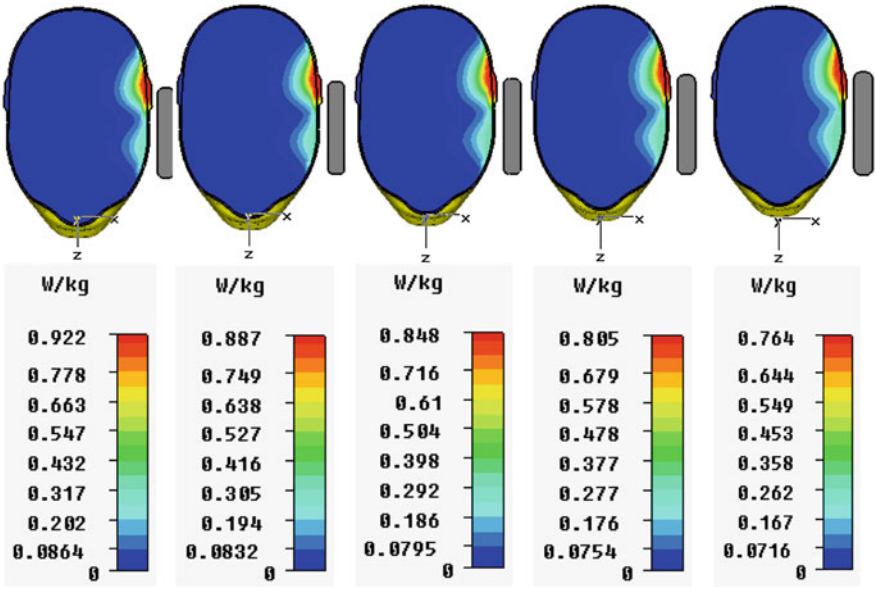


Fig. 10.6 Distributions of the (10 g) SAR at xz plane for 1800 MHz

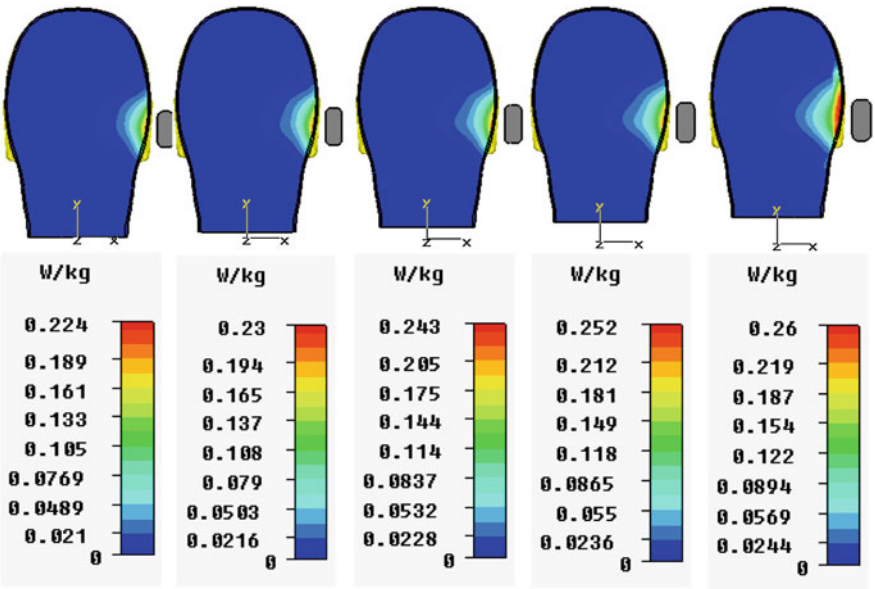


Fig. 10.7 The distributions of the (1 g) SAR at xy plane for 1800 MHz

1800 MHz, vary with the size of the human's head at each frequency. Also the sizes of the head have an effect on the mobile terminal antenna designer parameters, and this effect can't be eliminated, because it is an electromagnetic

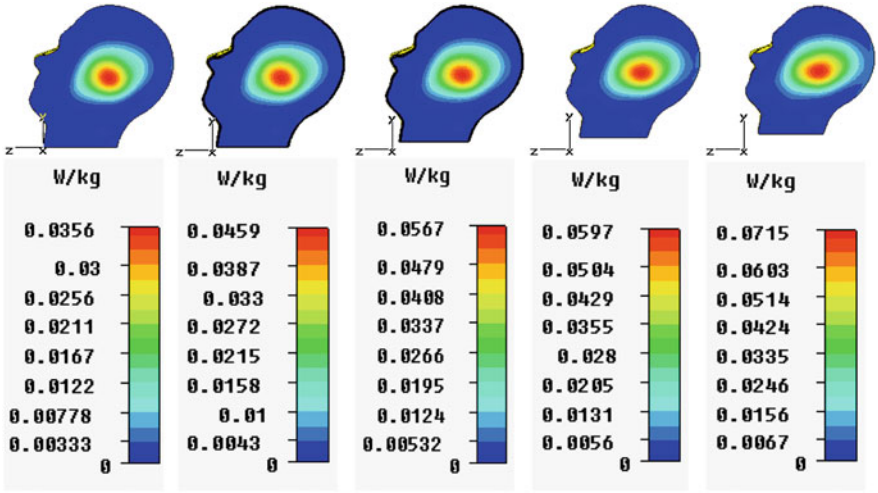


Fig. 10.8 Distributions of the local SAR at x = 0 plane for 900 MHz

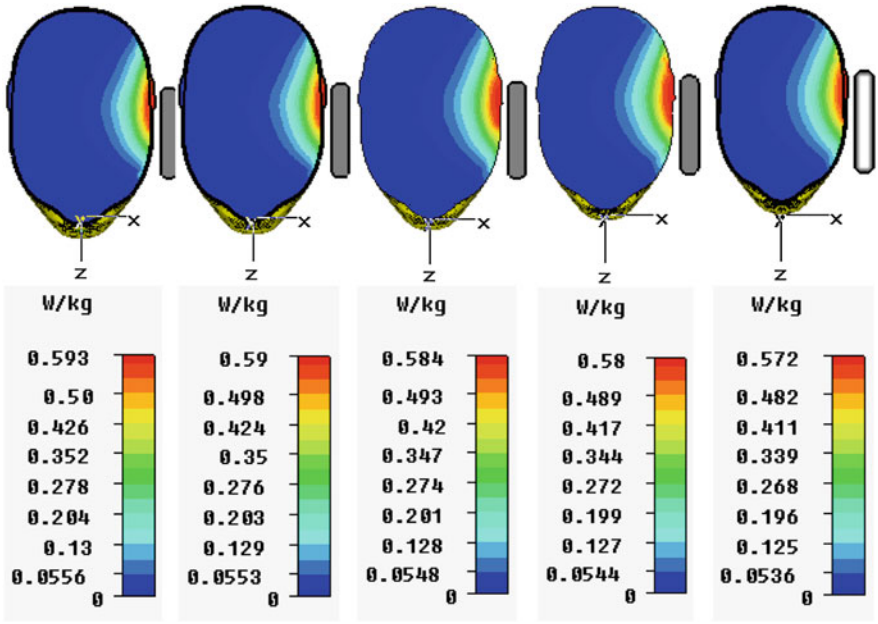


Fig. 10.9 Distributions of the (10 g) SAR at xz plane for 900 MHz

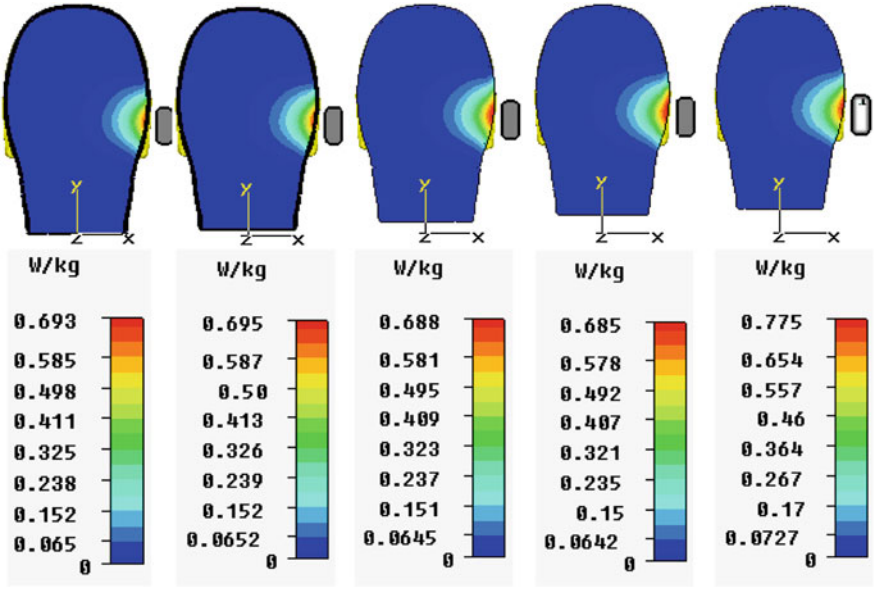


Fig. 10.10 Distributions of the (1 g) SAR at xy plane for 900 MHz

characteristic. The obtained results show that the spatial-peak SAR values as averaged over 1 g on the human head obtained with a radiated power of 0.125 W for all simulations are well below the limit of 1.6 W/kg, which is recommended by FCC and ICNIRP [12–14].

References

1. International Association of Engineers [Online]. Available: http://www.iaeng.org
2. El Dein AZ, Amr A (2010) Specific absorption rate (SAR) induced in human heads of various sizes when using a mobile phone at 900 and 1800 MHz. Lecture notes in engineering and computer science: Proceeding of the World Congress on Engineering 2010, Vol I, WCE 2010, 30 June–2 July, London, UK, pp 759–763
3. Kitchen R (2001) RF and microwave radiation safety handbook, Chapter 3, 2nd edn. Newnes, Oxford, pp 47–85
4. CST Microwave studio site. Available: http://www.cst.com/
5. Kiminami K, Iyama T, Onishi T, Uebayashi S (2008) Novel specific absorption rate (SAR) estimation method based on 2-D scanned electric fields. IEEE Trans Electromagn Compat 50(4):828–836
6. Watanabe S, Taki M, Nojima T, Fujiwara O (1996) Characteristics of the SAR distributions in a head exposed to electromagnetic fields radiated by a hand-held portable radio. IEEE Trans Microwave Theory Tech 44(10):1874–1883
7. Hadjem A, Lautru D, Dale C, Wong MF, Fouad-Hanna V, Wiart J (2004) Comparison of specific absorption rate (SAR) induced in child-sized and adult heads using a dual band mobile phone. Proceeding on IEEE MTT-S Int. Microwave Symposium Digest, June 2004

8. Kivekäs O, Ollikainen J, Lehtiniemi T, Vainikainen P (2004) Bandwidth, SAR, and efficiency of internal mobile phone antennas. IEEE Trans Electromagn Compat 46(1):71–86
9. Beard BB et al (2006) Comparisons of computed mobile phone induced SAR in the SAM phantom to that in anatomically correct models of the human head. IEEE Trans Electromagn Compat 48(2):397–407
10. Gabriel C (1996) Compilation of the Dielectric Properties of Body Tissues at RF and Microwave Frequencies."Brooks Air" Force Technical Report AL/OE-TR-1996-0037 [Online]. Available: http://www.fcc.gov/cgi-bin/dielec.sh
11. El Dein AZ (2010) Interaction between the human body and the mobile phone. Book Published by LAP Lambert Academic, ISBN 978-3-8433-5186-7
12. FCC, OET Bulletin 65, Evaluating Compliance with FCC Guidelines for Human Exposure to Radiofrequency Electromagnetic Fields. Edition 97-01, released December, 1997
13. IEEE C95.1-1991 (1992) IEEE standard for safety levels with respect to human exposure to radio frequency electromagnetic fields, 3 kHz to 300 GHz. Institute of Electrical and Electronics Engineers, Inc., New York
14. European Committee for Electrotechnical Standardization (CENELEC) (1995) Prestandard ENV 501 66-2, Human exposure to electromagnetic fields. High frequency (10 kHz to 300 GHz)

Chapter 11
A Medium Range Gbps FSO Link

Extended Field Performance Measurements

J. A. R. Pacheco de Carvalho, N. Marques, H. Veiga, C. F. Ribeiro Pacheco and A. D. Reis

Abstract Wireless communications have been increasingly important. Besides Wi-Fi, FSO plays a very relevant technological role in this context. Performance is essential, resulting in more reliable and efficient communications. A 1.14 km FSO medium range link has been successfully implemented for high requirement applications at Gbps. An extended experimental performance evaluation of this link has been carried out at OSI levels 1, 4 and 7, through a specifically planned field test arrangement. Several results, obtained namely from simultaneous measurements of powers received by the laser heads for TCP, UDP and FTP experiments, are presented and discussed.

J. A. R. P. de Carvalho (✉) · C. F. Ribeiro Pacheco · A. D. Reis
Unidade de Detecção Remota, Universidade da Beira Interior, 6201-001, Covilhã, Portugal
e-mail: pacheco@ubi.pt

C. F. Ribeiro Pacheco
e-mail: a17597@ubi.pt

A. D. Reis
e-mail: adreis@ubi.pt

N. Marques · H. Veiga
Centro de Informática, Universidade da Beira Interior, 6201-001, Covilhã, Portugal
e-mail: nmarques@ubi.pt

H. Veiga
e-mail: hveiga@ubi.pt

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_11,
© Springer Science+Business Media B.V. 2011

11.1 Introduction

Wi-Fi and FSO are wireless communications technologies whose importance and utilization have been growing for their versatility, mobility, speed and favourable prices.

Wi-Fi uses microwaves in the 2.4 and 5 GHz frequency bands and IEEE 802, 11a, b, g standards. Nominal transfer rates up to 11 (802.11b) and 54 Mbps (802.11a, g) are specified [1]. It has been used in ad hoc and infrastructure modes. Point-to-point and point-to-multipoint configurations are used both indoors and outdoors, requiring specific directional and omnidirectional antennas. FSO uses laser technology to provide point-to-point communications e.g. to interconnect LANs of two buildings having line-of-sight. FSO was developed in the 1960s for military and other purposes, including high requirement applications. At present, speeds typically up to 2.5 Gbps are possible and ranges up to a few km, depending on technology and atmospheric conditions. Interfaces such as fast Ethernet and Gigabit Ethernet are used to communicate with LAN's. Typical laser wavelengths of 785, 850 and 1550 nm are used. In a FSO link the transmitters deliver high power light which, after travelling through atmosphere, appears as low power light at the receiver. The link margin of the connection represents the amount of light received by a terminal over the minimum value required to keep the link active: $(\text{link margin})_{\text{dB}} = 10 \log_{10} (P/P_{min})$, where P and P_{min} are the corresponding power values, respectively.

There are several factors related to performance degradation in the design of a FSO link: distance between optical emitters; line of sight; alignment of optical emitters; stability of the mounting points; atmospheric conditions; water vapour or hot air; strong electromagnetic interference; wavelength of the laser light [2]. A redundant microwave link is always essential, as the laser link can fail under adverse conditions and communications are interrupted. Several studies and implementations of FSO have been reported [3, 4]. FSO has been used in hybrid systems for temporary multimedia applications [5].

Performance has been a very important issue, resulting in more reliable and efficient communications. Telematic applications have specific performance requirements, depending on application. New telematic applications present special sensitivities to performances, when compared to traditional applications. E.g. requirements have been quoted as: for video on demand/moving images, 1–10 ms jitter and 1–10 Mbps throughput; for Hi Fi stereo audio, jitter less than 1 ms and 0.1–1 Mbps throughputs [6].

Several performance measurements have been made for Wi-Fi [7, 8]. FSO and fiber optics have been applied at the University of Beira Interior Campus, at Covilhã City, Portugal, to improve communications quality [9–12]. In the present work we have further investigated that FSO link, for extended performance evaluation at OSI levels 1, 4 and 7.

The rest of the paper is structured as follows: Chap. 2 presents the experimental details i.e. the measurement setup and procedure. Results and discussion are presented in Chap. 3. Conclusions are drawn in Chap. 4.

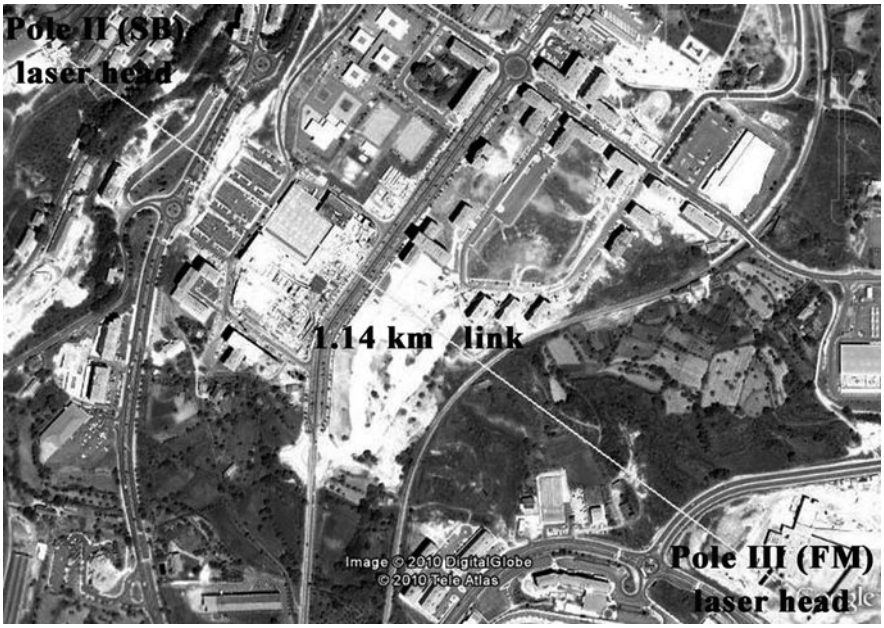


Fig. 11.1 View of the 1.14 km laser link between Pole II (SB) and Pole III (FM)

11.2 Experimental Details

The main experimental details, for testing the quality of the FSO link, are as follows.

A 1 Gbps full-duplex link was planned and implemented, to interconnect the LAN at the Faculty of Medicine building and the main University network, to support medical imaging, VoIP, audio and video traffics [9, 10]. Then, a FSO laser link at 1 Gbps full-duplex, over a distance of 1.14 km, was created to interconnect the Faculty of Medicine (FM) building at Pole III and the Sports (SB) building at Pole II of the University (Fig. 11.1).

We have chosen laser heads from FSONA (Fig. 11.2) to implement the laser link at a laser wavelength of $\lambda = 1550$ nm for eye safety, where allowable laser power is about fifty times higher at 1550 nm than at 800 nm [2–13]. Each laser head comprised two independent transmitters, for redundancy, and one wide aperture receiver. Each laser had 140 mW of power, resulting in an output power of 280 mW (24.5 dBm). 1000-Base-LX links over OM3 50/125 μm fiber were used to connect the laser heads to the LANs.

For a matter of redundancy a 802.16d WiMAX point-to-point link at 5.4 GHz was available, where data rates up to either 75 Mbps or 108 Mbps were possible in normal mode or in turbo mode, respectively [14]. This link was used as a backup link for FM-SB communications, through configuration of two static routing entries in the switching/routing equipment [9].

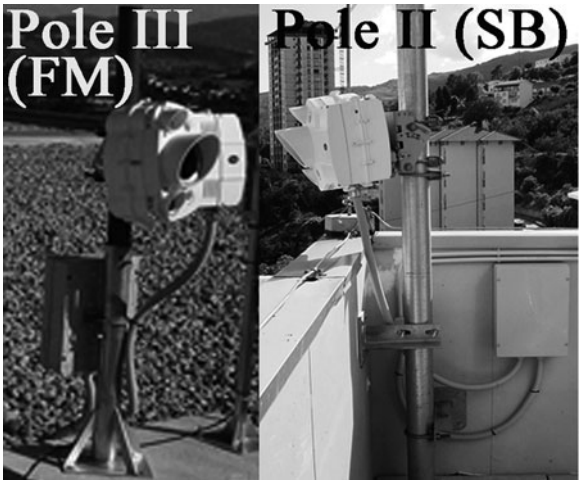


Fig. 11.2 View of the laser heads at FM (Pole III) and SB (Pole II)

Performance tests of the FSO link were made under favourable weather conditions. During the tests we used a data rate mode for the laser heads which was compatible with Gigabit Ethernet. At OSI level 1 (physical layer), received powers were simultaneously measured for both laser heads. Data were collected. from the internal logs of the laser heads, using STC (SONAbeam Terminal Controller) management software [13]. At OSI level 4 (transport layer), measurements were made for TCP connections and UDP communications using Iperf software [15], permitting network performance results to be recorded. Both TCP and UDP are transport protocols. TCP is connection-oriented. UDP is connectionless, as it sends data without ever establishing a connection. For a TCP connection over a link, TCP throughput was obtained. For a UDP communication, we obtained UDP throughput, jitter and percentage loss of datagrams. TCP packets and UDP datagrams of 1470 bytes size were used. A window size of 8 kbytes and a buffer size of the same value were used for TCP and UDP, respectively.

A specific field test arrangement was planned and implemented for the measurements (Fig. 11.3). Two PC's having IP addresses 192.168.0.2 and 192.168.0.1 were setup as the Iperf server and client, respectively. The PCs were HP computers, with 3.0 GHz Pentium IV CPUs, running Windows XP. The server had a better RAM configuration than the client. They were both equipped with 1000Base-T network adapters. Each PC was connected via 1000Base-T to a C2 Enterasys switch [16]. Each switch had a 1000Base-LX interface. Each interface was intended to establish a FSO link through two laser heads, as represented in Fig. 11.3. The laser heads were located at Pole II and Pole III, at the SB and FM buildings, respectively. The experimental arrangement could be remotely accessed through the FM LAN. In the UDP tests a bandwidth parameter of 300 Mbps was used in the Iperf client. Jitter, which represents the smooth mean of differences between consecutive transit times, was continuously computed by the server, as specified by RTP in RFC 1889 [17]. RTP provides end-to-end network transport functions appropriate for applications

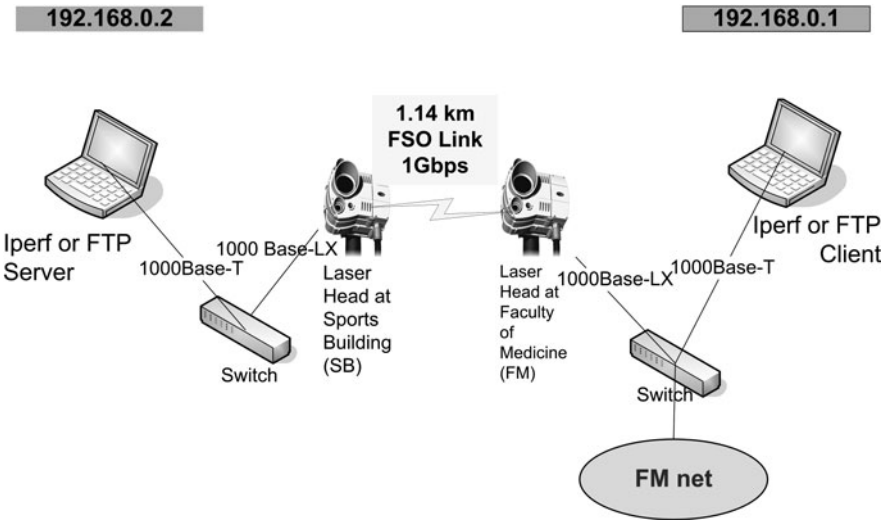


Fig. 11.3 Field tests setup scheme for the FSO link

transmitting real-time data, e.g. audio, video, over multicast or unicast network services. At OSI level 7 (application layer) the setup given in Fig. 11.3 was also used for measurements of FTP transfer rates through FTP server and client applications installed in the PCs. Each measurement corresponded to a single FTP transfer, using a 2.71 Gbyte file. Whenever a measurement was made at either OSI level 4 or 7, data were simultaneously collected at OSI level 1. Batch command files were written to enable the TCP, UDP and FTP tests. The results, obtained in batch mode, were recorded as data files in the client PC disk.

11.3 Results and Discussion

Several sets of data were collected and processed. The TCP, UDP and FTP experiments were not simultaneous. The corresponding results are shown for TCP in Fig. 11.4, for UDP in Fig. 11.6 and FTP in Fig. 11.8. The average received powers for the SB and FM laser heads, mostly ranged high values in the 25–35 μW interval which corresponds to link margins of 4.9–6.4 dB (considering Pmin = 8 μW). From Fig. 11.4 it follows that TCP average throughput (314 Mbps) is very steady; some small peaks arise for throughput deviation. Figure 11.5 illustrates details of TCP results over a small interval. Figure 11.6 shows that UDP average throughput (125 Mbps) is fairly steady, having a small steady throughput deviation. The jitter is small, usually less than 1 ms, while percentage datagram loss is practically negligible. Figure 11.7 illustrates details of UDP-jitter results over a small interval. Figure 11.8 shows that average FTP throughput (344 Mbps) is very steady, having low throughput deviation.

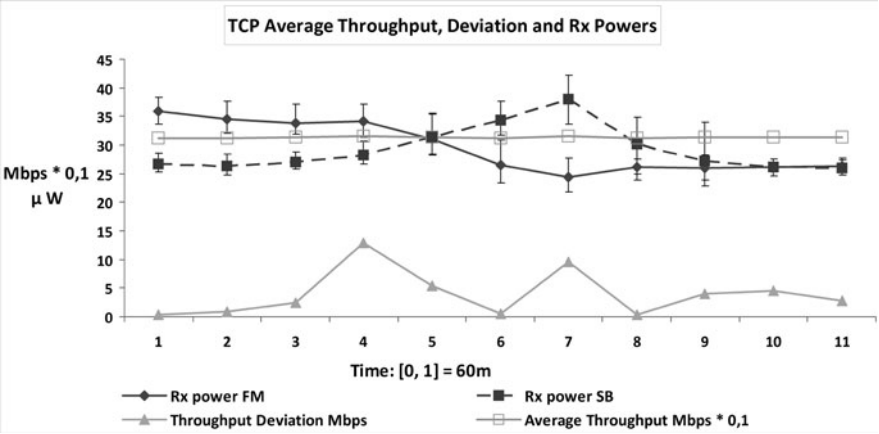


Fig. 11.4 TCP results

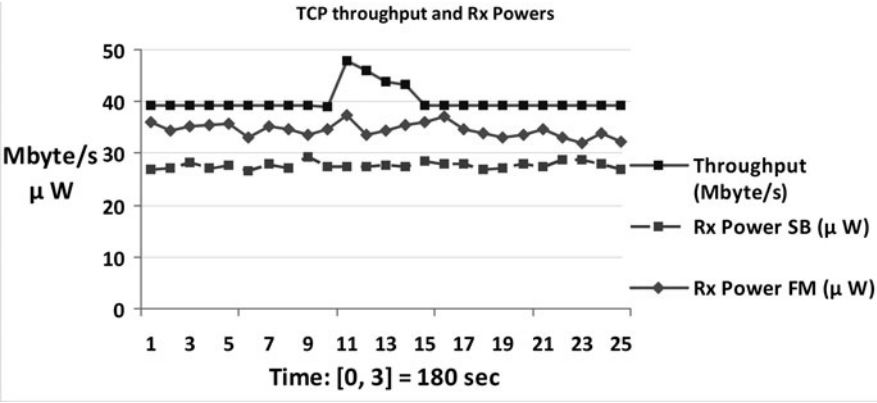


Fig. 11.5 Details of TCP results

Figure 11.9 illustrates details of FTP results over a small interval. Transfer rates of the PC's disks are always a limitation in this type of FTP experiments. In all cases, high values of average received powers were observed. The quantities under analysis did not show on average significant variations even when the received powers varied. The results here obtained complement previous work by the authors [9–12]. Generally, for our experimental conditions, the FSO link has exhibited very good performances at OSI levels 4 and 7.

Besides the present results, it must be mentioned that we have implemented a VoIP solution based on Cisco Call Manager [18]. VoIP, with G.711 and G729A coding algorithms, has been working over the laser link without any performance problems. Tools such as Cisco IP Communicator have been used. Video and sound have also been tested through the laser link, by using eyeBeam Softphone CounterPath software [19]. Applications using the link have been well-behaved.

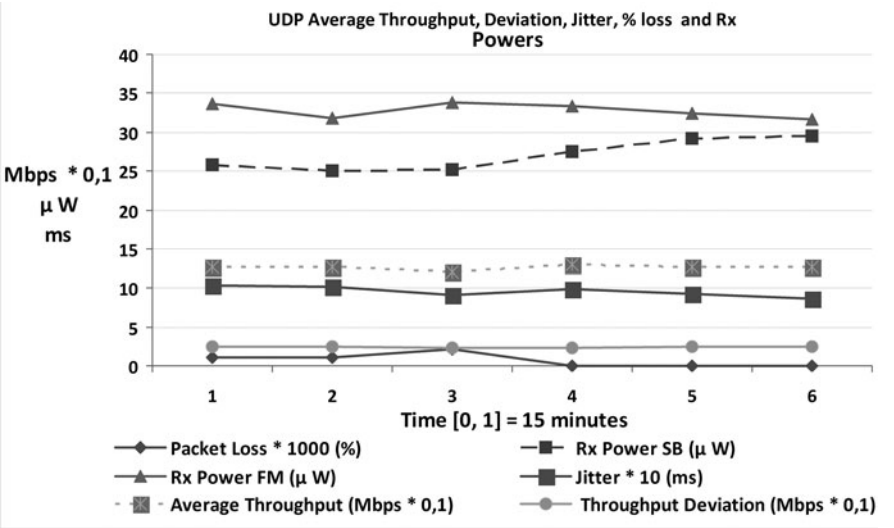


Fig. 11.6 UDP results; 300 Mbps bandwidth parameter

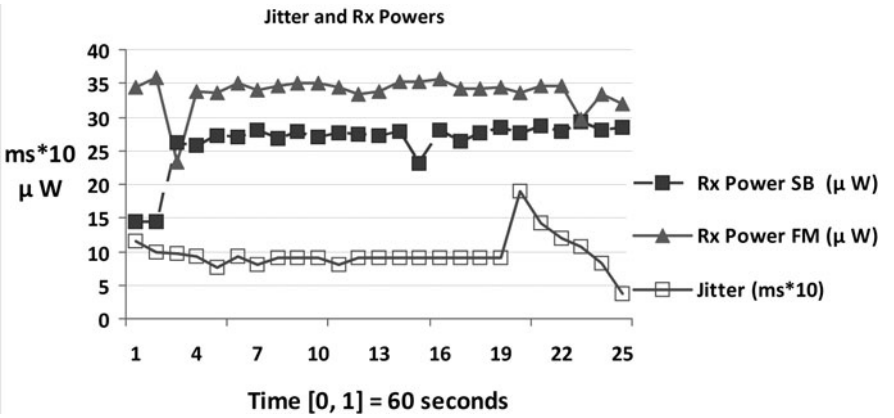


Fig. 11.7 Details of UDP-jitter results; 300 Mbps bandwidth parameter

11.4 Conclusions

A FSO laser link at 1 Gbps has been successfully implemented over 1.14 km along the city, for interconnecting Poles of the University and support high requirement applications.

A field test arrangement has been planned and implemented, permitting extended performance measurements of the FSO link at OSI levels 1, 4 and 7. At OSI level 1, received powers were simultaneously measured in both laser heads.

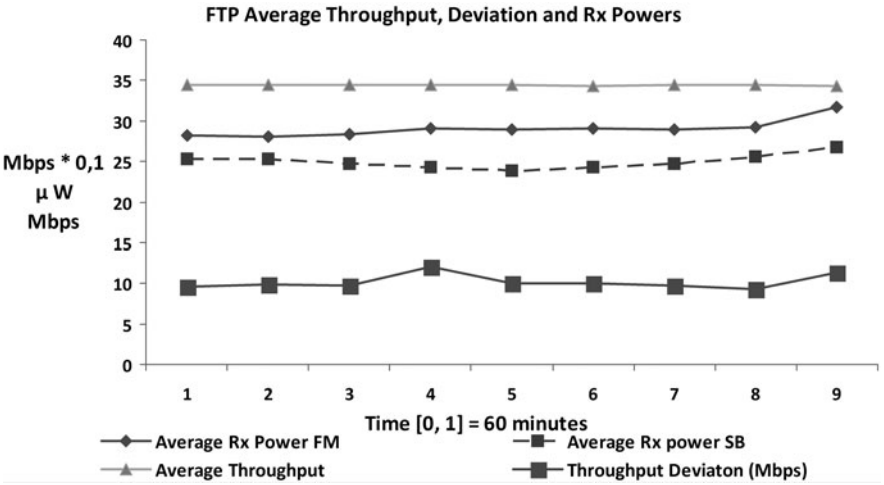


Fig. 11.8 FTP results

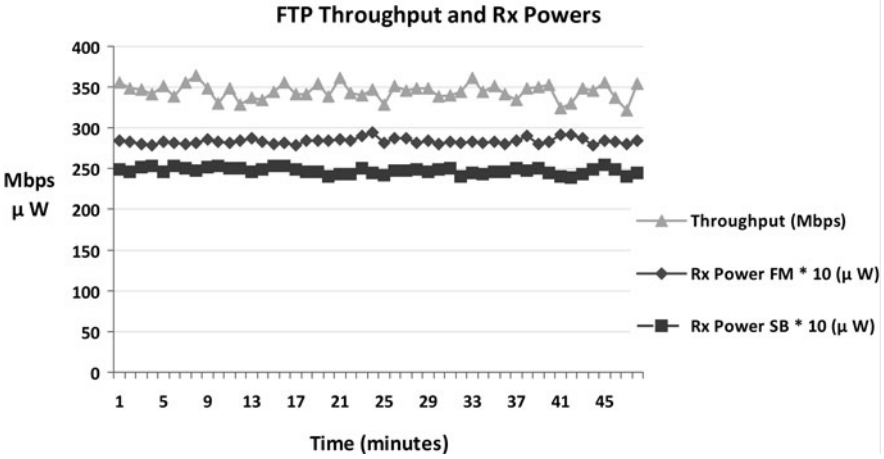


Fig. 11.9 Details of FTP results

At OSI level 4, TCP throughput, jitter and percentage datagram loss were measured. Through OSI level 7, FTP transfer rate data were acquired. Under favourable weather conditions, when the measurements were carried out, the link has behaved very well, giving very good performances. Applications such as VoIP, video and sound, have been well-behaved. Further measurements are planned under several experimental conditions.

Acknowledgments Supports from University of Beira Interior and FCT (Fundação para a Ciência e a Tecnologia)/POCI2010 (Programa Operacional Ciência e Inovação) are acknowledged. We acknowledge Hewlett Packard and FSONA for their availability.

References

1. IEEE Std 802.11-2007 (2007) IEEE Standard for Local and metropolitan area networks-Specific Requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications (October 10, 2007); http://standards.ieee.org/getieee802
2. Rockwell DA, Mecherle GS (2001) Wavelength selection for optical wireless communication systems. Proc SPIE 4530:27–35
3. Amico MD, Leva A, Micheli B (2003) Free space optics communication systems: first results from a pilot field trial in the surrounding area of Milan Italy. IEEE Microwave Wirel Compon Lett 13(8):305–307 August
4. Löschnigg M, Mandl P, Leitgeb E (2009) Long-term performance observation of a free space optics link. In: Proceedings of the 10th International Conference on Telecommunications-Contel, Zagreb, Croatia, June 8–10, pp 305–310
5. Mandl P, Chlestil Ch, Zettl K, Leitgeb E (2007) Hybrid systems using optical wireless, fiber optics and WLAN for temporary multimedia applications. In: Proceedings of the 9th International Conference on Telecommunications-Contel, Zagreb, Croatia, June 13–15, pp 73–76
6. Monteiro E, Boavida F (2002) Engineering of informatics networks, 4th edn. FCA-Editor of Informatics Ld, Lisbon
7. Pacheco de Carvalho JAR, Gomes PAJ, Veiga H, Reis AD (2008) Development of a university networking project. In: Putnik GD, Manuela Cunha M (eds) Encyclopedia of Networked and Virtual Organizations. IGI Global, Hershey, pp 409–422
8. Pacheco de Carvalho JAR, Veiga H, Marques N, Ribeiro Pacheco CF, Reis AD (2010) Laboratory performance of Wi-Fi WEP point-to-point links: a case study. Lecture notes in engineering and computer science: Proceedings of The World Congress on Engineering, WCE 2010, vol I, London, UK, 30 June–2 July, pp 764–767
9. Pacheco de Carvalho JAR, Gomes PAJ, Veiga H, Reis AD (2007) Wi-Fi and very high speed optical links for data and voice communications. In: Proc. 2ª Conferência Ibérica de Sistemas e Tecnologias de Informação, Universidade Fernando Pessoa, Porto, Portugal, 21–23 June, pp 441–452
10. Pacheco de Carvalho JAR, Veiga H, Gomes PAJ, Reis AD (2008) Experimental performance evaluation of a very high speed free space optics link at the university of Beira interior campus: a case study. In: Proc. SEONs 2008- VI Symposium on Enabling Optical Network and Sensors Porto, Portugal, 20–20 June, pp 131–132
11. Pacheco de Carvalho JAR, Veiga H, Gomes PAJ, Ribeiro Pacheco CFFP, Reis AD (2008) Experimental performance study of a very high speed free space optics link at the university of Beira interior campus: a case study. In: Proc. ISSPIT 2008-8th IEEE International Symposium on Signal Processing and Information Technology Sarajevo. Bosnia and Herzegovina, December 16–19, pp 154–157
12. Pacheco de Carvalho JAR, Marques N, Veiga H, Ribeiro Pacheco CF, Reis AD (2010) Field performance measurements of a Gbps FSO link at Covilha City, Portugal. Lecture notes in engineering and computer science: Proceedings of the world congress on engineering, WCE 2010, Vol I, 30 June–2 July, London, UK, pp 814–818
13. Web site http://www.fsona.com; SONAbeam 1250-S technical data; SONAbeam Terminal Controller management software
14. Web site http://www.alvarion.com; Breeze NET B100 data sheet
15. Web site http://dast.nlanr.net; Iperf software
16. Web site http://www.enterasys.com; C2 switch technical manual
17. Network Working Group. RFC 1889-RTP: A Transport Protocol for Real Time Applications, http://www.rfc-archive.org
18. Web site http://www.cisco.com; Cisco Call Manager; Cisco IP Communicator
19. Web site http://www.counterpath.com; eyeBeam Softphone CounterPath software

Chapter 12
A Multi-Classifier Approach for WiFi-Based Positioning System

Jikang Shin, Suk Hoon Jung, Giwan Yoon and Dongsoo Han

Abstract WLAN fingerprint-based positioning systems are a viable solution for estimating the location of mobile stations. Recently, various machine learning techniques have been applied to the WLAN fingerprint-based positioning systems to further enhance their accuracy. Due to the noisy characteristics of RF signals as well as the lack of the study on environmental factors affecting the signal propagation, however, the accuracy of the previously suggested systems seems to have a strong dependence on numerous environmental conditions. In this work, we have developed a multi-classifier for the WLAN fingerprint-based positioning systems employing a combining rule. According to the experiments of the multi-classifier performed in various environments, the combination of the multiple numbers of classifiers could significantly mitigate the environment-dependent characteristics of the classifiers. The performance of the multi-classifier was found to be superior to that of the other single classifiers in all test environments; the average error distances and their standard deviations were much more improved by the multi-classifier in all test environments.

J. Shin (✉) · S. H. Jung
Department of Information and Communications Engineering, Korea Advanced Institute of Science and Technology, 373-1 Kusong-Dong, Yuseong-gu, Daejeon, 305-701, Korea
e-mail: scrash@kaist.ac.kr

S. H. Jung
e-mail: sh.jung@kaist.ac.kr

G. Yoon
Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, 373-1 Kusong-Dong, Yuseong-gu, Daejeon, 305-701, Korea
e-mail: gwyoon@ee.kaist.ac.kr

D. Han
Department of Computer Science, Korea Advanced Institute of Science and Technology, 373-1 Kusong-Dong, Yuseong-gu, Daejeon, 305-701, Korea
e-mail: dshan@kaist.ac.kr

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_12,
© Springer Science+Business Media B.V. 2011

12.1 Introduction

With the explosive proliferation of smart phones, WLAN (Wireless Local Area Network)-based positioning systems have increasingly become a main stream in Location-based Service (LBS) regimes. Compared with other technologies such as GPS [1], RFID [2], GSM [3], Ultrasonic [4], infrared-based systems [5], etc., the WLAN-based positioning systems have some advantages in terms of coverage and costs. Most of the researches on the WLAN-based positioning systems have used the so-called Received Signal Strength Indication (RSSI) from the wireless network access points mainly because the RSSI (or called fingerprint) is relatively easy to obtain using software and also one of the most relevant factors for positioning.

Some studies have been reported that consider other factors such as Signal to Noise Ratio (SNR), Angle of Arrival (AOA), and Time of Arrival (TOA) for positioning systems. Milos et al. [6] examined the SNR as an additional input factor and reported that the consideration of both SNR and RSSI could increase the performance of the WLAN-based positioning system. Yamasaki et al. [7] reported that the AOA and TOA are also important factors in positioning. However, the acquisition of the factors including the AOA, TOA, and SNR are not always possible in every wireless network interface cards. Thus, the RSSI appears to have been adopted as a primary factor for the WLAN-based positioning systems.

In fact, utilizing the strengths of Radio Frequency (RF) signals for the positioning may not be a simple work. Due to the intrinsic characteristics of the RF signals like multipath fading and interference between signals, the signal strength may severely change depending on the materials used, the positions of doors and windows, the widths of the passages, the numbers of APs deployed, etc. Even if the fundamental parameters are known previously, the derivation of the path loss function of a WLAN signal is extremely complex. In this reason, the WLAN fingerprint-based positioning systems have mostly used to take statistical approaches [6].

The statistical approaches previously suggested have applied various machine learning techniques to derive the positions from the measured fingerprints [2, 8–15]. Those techniques usually are comprised of two phases: off-line and on-line phases. In the off-line phase, fingerprints are captured at various positions of target place and stored in a database called a radio-map. In the on-line phase, the location of a fingerprint is estimated by comparing it with the stored fingerprints in the database.

The main problem of the WLAN fingerprint-based positioning systems is that the system performance is too much environment-dependent; in other words, there are not yet any general solutions available for the WLAN fingerprint-based positioning systems. Each system is designed to tackle different environments, and there is no analysis on the relation between the algorithm used and the test

environments. One method may outperform other methods in an environment, but it may show inferior results in other environments. For instance, Youssef et al. [12] suggested a joint-clustering technique and confirmed in their evaluation that their proposed algorithm outperformed RADAR [2]. According to the experiment by Wilson et al. [11], however, the RADAR was found to have a superior performance as compared to the joint-clustering technique. Similarly, this kind of problem was also observed in our experiments.

In this paper, we introduce a multi-classifier for the application of the WLAN fingerprint-based positioning systems. We have combined multiple classifiers to become an efficient environment-independent classifier that can realize the more stable and higher estimation accuracy in a variety of the environments. The motivation for using a multiple number of classifiers lies in the fact that the classifier performance is severely environment-dependent; thus, if we can select the most accurate classifier for a given situation, we may be able to achieve even better performance in diverse environments.

In this work, a multiple number of classifiers were combined using the Bayesian combination rule [16] and majority vote [17]. To prove the combination effects of the classifiers, we have evaluated the proposed system in three different environments. The evaluation results revealed that the multi-classifier could outperform the single classifiers in terms of the average error distances and their standard deviations. This indicates that the proposed combining method is much more effective in mitigating the environment-sensitive characteristics of the WLAN-based positioning systems.

The remainder of this paper is organized as follows. The overview on the WLAN fingerprint-based positioning is given in Sect. 12.2. We introduce a multi-classifier for the WLAN fingerprint-based positioning systems in Sect. 12.3. Section 12.4 describes the experiment setup and results. Section 12.5 summarizes this work and suggests the future work.

12.2 Related Work

The location estimation using the so-called WLAN fingerprint often refers to the machine learning problem due to the high complexity of the signal propagation estimation. In this reason, various machine learning techniques have been applied. The RADAR system developed by Bahl et al. [2] is considered one of the most representative WLAN fingerprint-based systems. In this system, the authors used the Pentium-based PCs as access points and also the laptop computers as mobile devices. The system uses the nearest neighbor heuristics and triangulation methods to infer a user location. It maintains the radio map which can chart the strength of the signals received from the different access points at some selected locations. Each signal-strength measurement is then compared against the radio map, and

then the best matching positions are averaged, enabling the location estimation. Roos et al. [10] proposed the probability-based system which uses the received signal strength samples to create the probability distributions of the signal strength for some known locations. Once an input instance is given, it matches to these probability distributions to find out the location of the mobile device with the highest probability. The histogram method suggested by Castro et al. [18] is another example of the probability-based system. Instead of using Gaussian distribution, it derives the distribution of the signal strength from the learning data. In addition, the adaptive neural networks [13], decision tree [14, 15], and support vector machine [19] are popular on the WLAN-based positioning systems; Kushki et al. [8] suggested the kernelized distance calculation algorithm for the inference of the location of the measured RSSI.

Recently, some researchers have focused on compensating the characteristics of the RF signals. Berna et al. [20] suggested the system using the database by considering the unstable factors related to open/close doors and humidity changing environments. They utilized some sensors to capture the current status of the environment. Yin [15] introduced the learning approach based on the temporally updated database in accordance with the current environment situation. Moraes [21] investigated the dynamic RSS mapping architecture. By Wilson Yeung et al. [11], the use of the RSSI was suggested that are transmitted from the mobile devices as an additional input. Thus, there are two types of databases: the RSSI transmitted by APs and the RSSI transmitted by mobile devices. In the on-line phase, the system inferences the multiple results from the databases and makes the final decision using the combining method.

Some research efforts [12, 22] have tackled the issue on how to reduce the computational overhead mainly because the client devices are usually small, self-maintained and stand-alone, having a significant limitation in their power supply. Youssef et al. [12] developed a joint-clustering technique for grouping some locations in order to reduce the computational cost of the system. In this method, a cluster is defined as a set of locations sharing the same set of access points. The location determination process is as follows: for a given RSSI data set, the strongest access points are selectively used to determine one cluster to search the most probable location. Chen et al. [22] suggested the method which selects the most discriminative APs in order to minimize the AP numbers used in the positioning system. This approach selects an appropriate subset of the existing features to the computational complexity problem. Reducing the number of APs is referred to as the dimension reduction in a signal space, which in turn reduces the computational overheads required on the mobile devices.

The weak spot of the WLAN fingerprint-based positioning systems is that their performance is severely environment-dependent. One system may outperform the other methods in an environment; it may show an inferior performance in other environments. To solve this problem, we suggest a multi-classifier approach for the application of the WLAN fingerprint-based positioning systems, leading to the more accurate results.

12.3 Proposed Method

We utilize the multiple numbers of classifiers using different algorithms to build a possibly environment-independent classifier [23]. The work of combining multiple numbers of classifiers to create a strong classifier has been a well-established research, particularly in the pattern recognition area, the so-called Multiple Classifier System (MCS) [24]. When it comes to the term "combining", it indicates a processing of selecting the most trustable prediction results attained from the classifiers.

At least, two reasons may justify the necessity of combining multiple classifiers [25]. First, there are a number of classification algorithms available that were developed from different theories and methodologies for the current pattern recognition applications. For a specific application problem, usually, each one of these classifiers could reach a certain degree of success, but maybe none of them is totally perfect or at least one of them is not so good as expected in practical applications. Second, for a specific recognition problem, there are often many types of features which could be used to represent and recognize some specific patterns. These features are also represented in various diversified forms and it is relatively hard to lump them together for one single classifier to make a decision. As a result, the multiple classifiers are needed to deal with the different features. It also results in a general problem on how to combine those classifiers with different features to yield the improved performance.

The location estimation using the WLAN fingerprint often refers to the classification problem because of the noisy characteristics of the RF signals. Many algorithms have been proposed based on the different machine learning techniques, but none of them could achieve the best performance in very diverse environments. At this point, we realized that utilizing the multiple numbers of classifiers could be a promising solution, as a general solution for the WLAN fingerprint-based positioning systems.

In this work, we combined the Bayesian combination rule [16] and majority vote [17] for our multi-classifier. The Bayesian combination rule gives weights to the decisions of classifier based on the information in a basis prepared in learning phase. Usually, the basis is given in a form of matrix called a confusion matrix. The confusion matrix is constructed by the cross-validation with learning data in the off-line phase. The majority vote is a simple algorithm, which chooses the one selected by more than a half of the classifiers.

Figure 12.1 illustrates the idea of our proposed system. In the off-line phase, the fingerprints are collected over the target environment as learning data. The fingerprint is a collection of the pair-wise data containing the MAC address of an access point and its signal strength. Usually, in one fingerprint, there are multiple tuples of this pair-wise data such as $\{<\mathrm{ap}_1,\ \mathrm{bssi}_1>, <\mathrm{ap}_2,\ \mathrm{bssi}_2>, <\mathrm{ap}_3,\ \mathrm{bssi}_3> \ldots.\}$. After attaching the collected location labels to the fingerprints, the database stores the labeled-fingerprint data.

After collecting the learning data, each classifier C constructs their own confusion matrix **M** (Fig. 12.2) using the cross-validation with the learning data. The

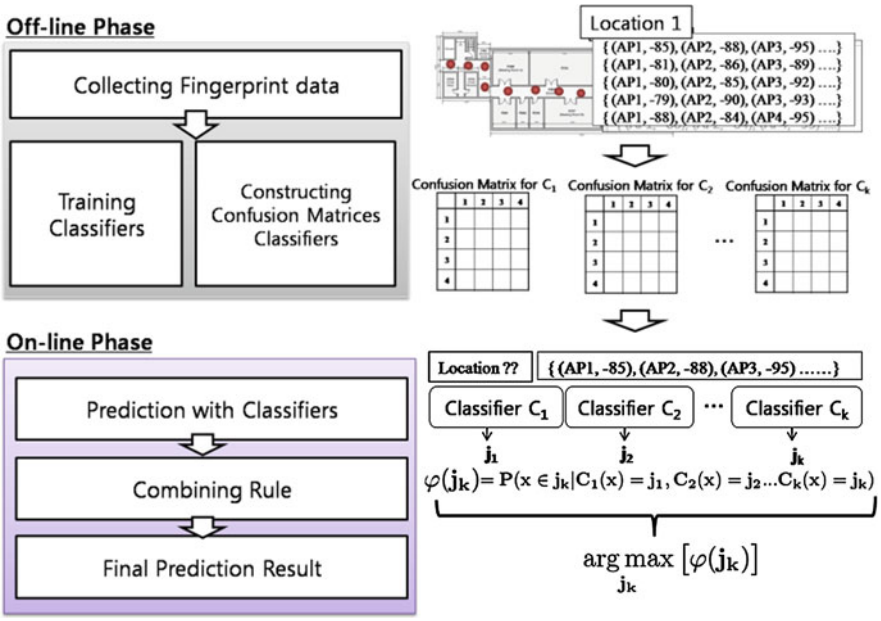

Off-line Phase
Collecting Fingerprint data
Training Classifiers
Constructing Confusion Matrices Classifiers
Location 1
{(AP1, -85), (AP2, -88), (AP3, -95) ....}
{(AP1, -81), (AP2, -86), (AP3, -89) ....}
{(AP1, -80), (AP2, -85), (AP3, -92) ....}
{(AP1, -79), (AP2, -90), (AP3, -93) ....}
{(AP1, -88), (AP2, -84), (AP4, -95) ....}
Confusion Matrix for C1
Confusion Matrix for C2
Confusion Matrix for Ck
On-line Phase
Prediction with Classifiers
Combining Rule
Final Prediction Result
Location ??
{(AP1, -85), (AP2, -88), (AP3, -95) .....}
Classifier C1
Classifier C2
Classifier Ck
φ(jk)= P(x ∈ jk|C1(x) = j1, C2(x) = j2...Ck(x) = jk)
arg max [φ(jk)]

Fig. 12.1 The overview of multi-classifier

		Actual Location																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	⋯	sum
Predicted Location	1	22	0	13	5	0	0	0	0	1	0	0	0	0	0	0	⋯	41
	2	3	6	17	3	0	0	1	0	0	0	0	0	0	0	0	⋯	30
	3	0	1	6	1	0	0	0	0	0	0	0	0	0	0	0	⋯	8
	4	18	14	9	6	5	12	0	7	2	1	11	1	0	0	2	⋯	96
	5	1	0	1	0	6	8	2	0	9	7	0	0	1	1	1	⋯	49
	6	5	23	2	32	32	25	1	12	0	21	6	1	0	0	1	⋯	231
	7	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	⋯	9
	8	0	0	0	0	0	0	4	23	4	0	0	1	8	1	9	⋯	64
	9	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	⋯	23
	10	0	1	0	0	0	0	13	0	1	16	6	10	15	8	10	⋯	91
	11	0	0	0	0	2	2	1	4	1	2	20	1	8	11	7	⋯	86
	12	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	⋯	12
	13	0	0	0	0	0	0	0	0	0	0	0	2	6	0	0	⋯	8
	14	0	0	0	0	0	0	5	0	6	0	0	20	2	18	0	⋯	54
	15	0	0	0	0	0	0	0	0	1	0	0	0	0	0	5	⋯	17
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋯
	sum	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	

Fig. 12.2 An example of confusion-matrix

confusion matrix would be used as an indicator of its classifier. If there are L possible locations in the positioning system, the $\mathbf{M}$ will be a $L \times L$ matrix in which the entry $\mathrm{M}_{i,j}$ denotes the number of the instances collected in *location i*, that is assigned as *location j* by the classifier.

From the matrix M, the total number of data collected in *location i* can be obtained as a row sum $\sum_{i=1}^{L} \mathrm{M}_{i,j}$, and the total number of data assigned to *location j* can be obtained as a column sum $\sum_{j=1}^{L} \mathrm{M}_{i,j}$ When there are K classifiers, there would be K confusion matrices $\mathrm{M}^{(k)}, 1 \leq k \leq k$.

In the on-line phase, for the measured Fingerprint x, the positioning results gained by K classifiers are $C_k(x) = j_k$, $1 \leq k \leq k$, and the j_k can be any location of the L possible locations. The probability that the decision made by the classifier C_k is correct can be measured as follows:

$$\varphi(j_k) = P(x \in j_k | C_1(x) = j_1, \ldots, C_k(x) = j_k) \tag{12.1}$$

Equation 12.1 is called the belief function, and the value of this function is called the belief value. Assuming that all classifiers are independent each other, and applying the Bayes' theorem to Eq. 12.1, the belief function $\varphi(j_k)$ can be reformulated as:

$$\varphi(j_k) = \prod_{\mathrm{i}=1}^{K} \frac{P(x \in j_k \cap C_i(x) = j_i)}{P(C_i(x) = j_i)} \tag{12.2}$$

The denominator and numerator in Eq. 12.2 can be calculated using the confusion matrix $\mathbf{M}$. The denominator indicates the probability that the classifier c_i will assign the unknown fingerprint x to j_i. This can be presented as follows:

$$P(C_i(x) = j_i) = \frac{\sum_{j=1}^{L} \mathrm{M}_{i,j}}{\sum_{i,j=1}^{L} \mathrm{M}_{i,j}} \tag{12.3}$$

The numerator in the Eq. 12.2 means the probability that the classifier c_i will assign the unknown fingerprint x collected in j_k to j_i. This term is simply described as below:

$$P(x \in j_k \cap C_i(x) = j_i) = \frac{\mathrm{M}_{j_k, j_i}}{\sum_{i,j=1}^{L} \mathrm{M}_{i,j}} \tag{12.4}$$

After applying Eqs. 12.3 and 12.4 to Eq. 12.2, Eq. 12.2 can be reformulated as:

$$\varphi(j_k) = \prod_{i=1}^{\mathrm{K}} \frac{\mathrm{M}_{j_k, j_i}}{\sum_{j=1}^{L} \mathrm{M}_{i,j}}$$

If more than a half of estimation of the classifiers pointed a specific location, the location would be selected as the final result. Otherwise, the belief value of each prediction is calculated, and the location with the highest belief value would be the final result. In case there are many locations with the same highest belief value, the

multi-classifier system determines the middle point of those locations as the final result.

For example, assume that there are three classifiers, *a*, *b*, and *c*, and there are three possible locations, *location 1*, *location 2* and *location 3*. After the off-line phase, the confusion matrices will be as follows:

$$\mathbf{M}^{(a)} = \begin{pmatrix} 18 & 4 & 7 \\ 2 & 12 & 3 \\ 0 & 4 & 10 \end{pmatrix}$$
$$\mathbf{M}^{(b)} = \begin{pmatrix} 12 & 6 & 6 \\ 3 & 9 & 3 \\ 2 & 5 & 11 \end{pmatrix}$$
$$\mathbf{M}^{(c)} = \begin{pmatrix} 14 & 2 & 2 \\ 4 & 11 & 5 \\ 2 & 7 & 13 \end{pmatrix}$$

If the classifiers *a*, *b*, and *c* assigned the unknown instance *x* to *location 1*, *location 2*, and *location 3*, respectively, the belief values of the predictions can be calculated as follows:

$$\varphi(j_a) = \frac{18}{29} \times \frac{3}{15} \times \frac{2}{22} = \frac{108}{9570}$$
$$\varphi(j_b) = \frac{4}{29} \times \frac{9}{15} \times \frac{7}{22} = \frac{252}{9570}$$
$$\varphi(j_c) = \frac{7}{29} \times \frac{3}{15} \times \frac{13}{22} = \frac{273}{9570}$$

The multi-classifier assigns the *location 3* to the unknown instance *x*, because the j_c, the prediction of the classifier *c*, has the highest belief value.

12.4 Evaluation

12.4.1 Experimental Setup

The performance of WLAN-based positioning systems depends on each environment itself where the evaluation is performed. In this reason, we evaluated the proposed multi-classifier in three different environments; Table 12.1 briefly illustrates the test environments. The testbed 1 implies an office environment; the dimension of the corridor in the office is 3 × 60 m. The office is on the third floor of the faculty building at the KAIST-ICC in Daejeon, South Korea. In the corridor, we have collected 100 samples of Fingerprints from 60 different locations. Each location is 1 m away from each other. The testbed 2 indicates another office

Table 12.1 Summary of testbeds

	Testbed 1	Testbed 2	Testbed 3
Type	Corridor	Corridor	Hall
Dimension (m)	3 × 60	4 × 45	15 × 15
Number of AP	60	45	25
Distance between RP (m)	1	1	3
Number of APs deployed	48	69	36
Avg. number of APs in one sample	16.6	16.8	13.9
Std.Dev of number of AP in sample	1.89	4.24	3.48

environment where the dimension of the corridor is 4 × 45 m. The office is located on the second floor of the Truth building at the KAIST-ICC. We have collected 100 samples of the Fingerprint from 45 different locations. Each location is 1 m away from each other. The testbed 3 implies a large and empty space inside the building located at the first floor of the Lecture building at the KAIST-ICC. The dimension of the space is 15 × 15 m. In the testbed 3, we have collected 100 samples of the Fingerprints from 25 different locations. Each location is 3 m away from each other.

Comparing the testbed 3 case with testbed 1 and 2 cases, there is no attenuation factors that may disturb any signal propagation. To collect the data, we have adopted the HTC-G1 mobile phone with Android 1.6 platform, and used the API provided by the platform. We have also used the half (50%) of the collected data as the learning data and the rest of data were used as the test data. To prove the better performance of the multi-classifier, we created the multi-classifier with three classifiers, k-NN (with $k = 3$) [2], Bayesian [9], and Histogram classifiers [10]; the performance of the multi-classifier was compared with these three classifiers, as shown in Table 12.2.

12.4.2 Results

We can observe from the results that none of the single classifier outperformed others in all three test environments. These results indicate that the performance of the WLAN fingerprint-based positioning systems is sensitively related to the environments and the multi-classifier is turned out to be much more effective in mitigating such characteristics of the WLAN signals.

Figure 12.3 reports the average error distance with respect to the different numbers of APs. From the Fig. 12.3a and b, the performances of the classifiers are quite different according to the test environments. Although the testbed 1 and testbed 2 look similar each other in indoor environments, the performances in testbed 1 are better than those in testbed 2. Especially, the average error distance of k-NN classifier in testbed 1 was 1.21 m when 15 APs were used for positioning,

Table 12.2 Summary of testbeds (meter)

		Avg	Std.Dev	Max	Min	90th Percentile
Testbed 1	*k*-NN	4.0	5.3	43	0	12.0
	Histogram	2.6	3.8	29	0	7.0
	Bayesian	2.8	3.9	25	0	7.0
	Multi	2.4	3.6	25	0	7.0
Testbed 2	*k*-NN	1.3	3.0	44	0	3.0
	Histogram	2.0	2.5	26	0	5.0
	Bayesian	1.3	1.8	17	0	3.0
	Multi	1.1	1.6	13	0	3.0
Testbed 3	*k*-NN	4.8	4.5	22.5	0	18.03
	Histogram	5.8	4.6	22.5	0	20.62
	Bayesian	5.6	5.1	22.5	0	21.21
	Multi	4.5	4.5	22.5	0	18.03

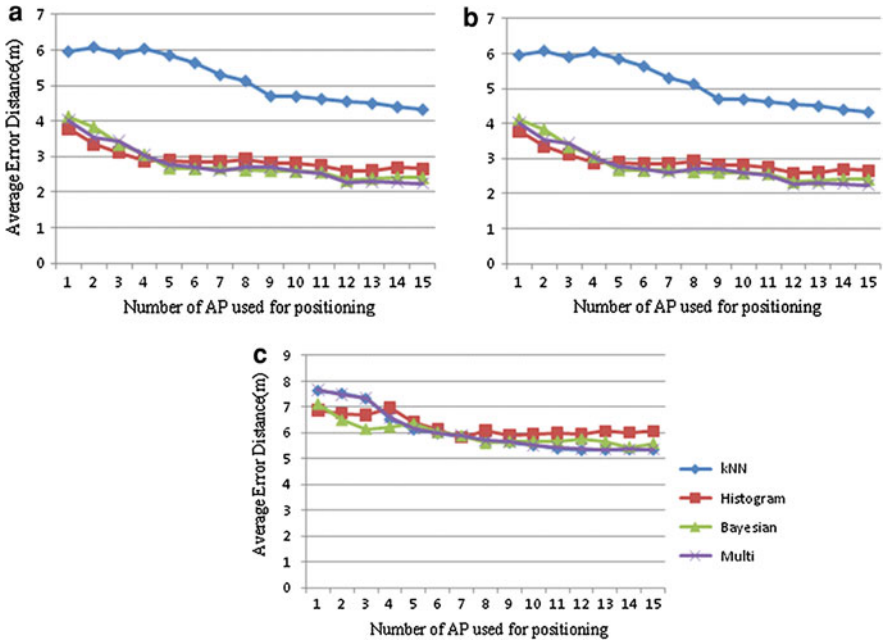


Fig. 12.3 Average error distance versus number of AP used for positioning in **a** Testbed 1, **b** Testbed 2, and **c** Testbed 3 respectively

while it was 4.6 m in testbed 2. In case of the histogram classifier, the average error distances were 1.9 and 2.7 m with 15 APs in testbed 1 and testbed 2, respectively. With the same condition, the Naïve Bayesian classifier's average error distances in the testbeds 1 and 2 were 1.25 and 2.47 m, respectively.

Compared with other classifiers, the multi-classifier showed the more improved results. In the testbeds 1 and 2, the average error distances of the multi-classifier

with 15 APs were 1.1 and 2.3 m, respectively. In the testbed 3, the accuracies of all classifiers are extremely poorer than the results in other testbeds. Based on the findings, it is believed that the WLAN fingerprint-based positioning systems can show better performance in the office environments as compared to the hall environments involving a few attenuation factors. As shown in the Fig. 12.3, the multi-classifier may clearly mitigate the environment-dependent characteristics of the single classifier. From the results shown in Fig. 12.3, we can conclude that the multi-classifier is effective for reducing error distance in localization.

Table 12.2 illustrates the performance summary of the classifiers. The standard deviation of the errors of the multi-classifier in the testbed 1 was 3.6 m, while the *k*-NN, Histogram, and Bayesian respectively showed 5.8, 3.8, and 3.9 m in their standard deviations. In the testbed 2, the standard deviations of the error of all classifiers were lower than the values in the testbed 1. The standard deviation of *k*-NN, Histogram and Bayesian were 3.0, 2.5, and 1.8 m, respectively. The standard deviation of the error of the *k*-NN, histogram, and Bayesian classifier in testbed 3 were 4.5, 4.6, and 5.1 m, respectively. These results confirm that the standard deviation of the errors of WLAN fingerprint-based positioning systems is also dependent on the environments. The proposed multi-classifier outperformed others in all testbeds in terms of the standard deviations of the error. In testbed 1, 2, and 3, the standard deviations of the errors of the multi-classifiers were 3.6, 1.6, and 4.5, respectively, which are higher or equivalent performance compared with others.

From the results, we confirmed that multi-classifier could mitigate the environment-dependent characteristics of the single classifier, and the performance of the multi-classifier was better than the others in all environments. Even if the improvement of performance was not remarkable, the results indicate that combining a number of classifiers is one of the promising approaches in constructing reliable and accurate WLAN fingerprint-based positioning systems.

12.5 Summary and Future Work

In this paper, we have presented an environment-independent multi-classifier for the WLAN fingerprint-based positioning systems in an effort to mitigate the undesirable environmental effects and factors. We have developed a combining method of the multiple numbers of classifiers for the purpose of the error-correction. For example, if a single classifier predicted wrong, the other classifiers correct it. In other words, the classifiers in the multi-classifier can complement each other.

We have evaluated the multi-classifier in three different environments with various environmental factors: the numbers of APs, the widths of corridor, the materials used, etc. The multi-classifier was constructed with three different classifiers; *k*-NN (with $k = 3$), Bayesian, and Histogram classifiers. As a result, the multi-classifier showed a consistent performance in the diverse test environments while other classifiers showed an inconsistent performance. The performance of

the multi-classifier tends to follow that of the single classifier showing the best performance. This means that the classifiers in the multi-classifier complement each other, and thus the errors are more effectively corrected.

For the next step, we are going to investigate a more efficient combining rule. In this work, we have mixed the Bayesian combining rule and majority vote; however, the performance enhancement was too marginal. Considering the complexity overhead of using the multiple numbers of classifiers, the multi-classifier may not be a cost-effective approach.

Finding the best combination of the classifiers will be another direction of our future work. We have tested only three classifiers, and two of them have taken similar approaches; the fingerprint is the only feature for positioning. There are numbers of systems considering various aspects of WLAN signals that use additional features. In the near future, we are going to implement and evaluate the multi-classifier with various types of classifiers.

Acknowledgments This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-(C1090-1011-0013)), and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2008-0061123).

References

1. Enge P, Misra P (1999) Special issue on global positioning system. Proc IEEE 87(1):3–15
2. Bahl P, Padmanabhan V (2000) RADAR: an in-building RF-based user location and tracking system. Proc IEEE Infocom 2:775–784
3. Drane C, Macnaughtan M, Scott C (1998) Positioning GSM telephones. IEEE Commun Mag 36(4):46–54
4. Priyantha N, Chakraborty A, Balakrishnan H (2000) The cricket location-support system. In: Proceedings of the 6th Annual International Conference on Mobile Computing and Networking, pp 32–43
5. Want R, Hopper A, Falcão V, Gibbons J (1992) The active badge location system. ACM Trans Inf Syst (TOIS) 10(1):102
6. Borenovic M, Neskovic A (2009) Comparative analysis of RSSI, SNR and noise level parameters applicability for WLAN positioning purposes. In: Proceedings of the IEEEEUROCON, pp 1895–1900
7. Yamasaki R, Ogino A, Tamaki T, Uta T, Matsuzawa N, Kato T (2005) TDOA location system for IEEE 802.11 b WLAN. In: Proceedings of IEEE. WCNC'05, pp 2338–2343
8. Kushki A, Plataniotis K, Venetsanopoulos A (2007) Kernel-based positioning in wireless local area networks. IEEE Trans Mobile Comput 6(6):689–705
9. Madigan D, Elnahrawy E, Martin R (2005) Bayesian indoor positioning systems. In: Proceedings of INFOCOM, pp 1217–1227
10. Roos T, Myllymaki P, Tirri H, Misikangas P, Sievanen J (2002) A probabilistic approach to WLAN user location estimation. Int J Wirel Inf Netw 9(3):155–164
11. Yeung W, Ng J (2007) Wireless LAN positioning based on received signal strength from mobile device and access points. In: IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, pp 131–137

12. Youssef M, Agrawala A, Shankar A (2003) WLAN location determination via clustering and probability distributions. In: Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, p 143
13. Borenovi M, Nekovic A, Budimir D (2009) Cascade-connected ANN structures for indoor WLAN positioning. Intell Data Eng Autom Learning-IDEAL 392–399
14. Chen Y, Yang Q, Yin J, Chai X (2006) Power-efficient access-point selection for indoor location estimation. IEEE Trans Knowl Data Eng 18(7):877–888
15. Yin J, Yang Q, Ni L (2008) Learning adaptive temporal radio maps for signal-strength-based location estimation. IEEE Trans Mobile Comput 7(7):869–883
16. Xu L, Krzyzak A, Suen C (1992) Methods of combining multiple classifiers and their application to hand writing recognition. IEEE Trans Syst Man Cybern 22:418–435
17. Kuncheva L (2001) Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recogn 34(2):299–314
18. Castro P, Chiu P, Kremenek T, Muntz R (2001) A probabilistic room location service for wireless networked environments. In: Proceeding of the 3rd International Conference on Ubiquitous Computing, pp 18–34
19. Brunato M, Battiti R (2005) Statistical learning theory for location fingerprinting in wireless LANs. Comput Netw 47(6):825–845
20. Berna M, Lisien B, Sellner B, Gordon G, Pfenning F, Thrun S (2003) A learning algorithm for localizing people based on wireless signal strength that uses labeled and unlabeled data. In: Proceedings of IJCAI, pp 1427–1428
21. Moraes L, Nunes B (2006) Calibration-free WLAN location system based on dynamic mapping of signal strength. In: Proceedings of the 4th ACM International Workshop on Mobility Management and Wireless Access, pp 92–99
22. Chen Y, Yin J, Chai X, Yang Q (2006) Power efficient access-point selection for indoor location estimation. IEEE Trans Knowl Data Eng 1(18):878–888
23. Shin J, Han D (2010) Multi-classifier for WLAN fingerprint-based positioning system. Lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering, WCE 2010, 30 June–2 July, London, UK, pp 768–773
24. Kittler J (1998) Combining classifiers: a theoretical framework. Pattern Anal Appl 1(1):18–27
25. Chen K, Wang L, Chi H (1997) Method of combining multiple classifiers with different features and their applications to text-independent speaker identification. Int J Pattern Recognit Artif Intell 11(3):417–445

Chapter 13
Intensity Constrained Flat Kernel Image Filtering, a Scheme for Dual Domain Local Processing

Alexander A. Gutenev

Abstract A non-linear image filtering scheme is described. The scheme is inspired by the dual domain bilateral filter but owing to much simpler pixel weighting arrangement the computation of the result is much faster. The scheme relies on two principal assumptions: equal weight of all pixels within an isotropic kernel and a constraint imposed on the intensity of pixels within the kernel. The constraint is defined by the intensity of the central pixel under the kernel. Hence the name of the scheme: Intensity Constrained Flat Kernel (ICFK). Unlike the bilateral filter designed solely for the purpose of edge preserving smoothing, the ICFK scheme produces a variety of filters depending on the underlying processing function. This flexibility is demonstrated by examples of edge preserving noise suppression filter, contrast enhancement filter and adaptive image threshold operator. The latter classifies pixels depending on local average. The versatility of the operators already discovered suggests further potentials of the scheme.

13.1 Introduction

The initial stimulus for the development of the proposed scheme arose from the need for noise suppressing, edge preserving smoothing filter with a quasi real-time performance. The literature on edge preserving smoothing is plentiful. The most successful methods employ a dual domain approach: they define the operation result as function of "distances" in two domains, spatial and intensity. The "distances" are measured from a reference pixel of the input image. Well known

A. A. Gutenev (✉)
Retiarius Pty Ltd., P.O. Box 1606, Warriewood, NSW 2102, Australia
e-mail: agutenev@retiarius-au.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_13,
© Springer Science+Business Media B.V. 2011

examples are SUSAN [1] or, in more general form, the bilateral filter [2]. The main design purpose of these filtering schemes was the adaptation of level of smoothing to the amount of detail available within the neighborhood of the reference pixel. The application of such schemes ranges from adaptive noise suppression to creation of cartoon-like scenes from real world photographs [3]. The main weakness of the bilateral filter is its slow execution speed due to exponential weighting functions applied to the image pixels in both spatial and intensity domains. There is a range of publications describing the ways of improving the calculation speed of the bilateral filter [4–6]. In this paper we shall see that the simplification of weighting functions in both spatial and intensity domains not only increases the speed of computation without loosing the essence of edge preserving smoothing, but also suggests a filter generation scheme, versatile enough to produce operators beyond the original task of adaptive smoothing.

13.2 Intensity Constrained Flat Kernel Filtering Scheme

13.2.1 Intensity Constrained Flat Kernel Filter as a Simplification of the Bilateral Filter

The bilateral filter is considered here in the light of its original purpose: single pass application. The output of the bilateral filter [2] is given by the formula [5]

$$I_p^b = \frac{1}{W_p^b} \sum_{q \in S} G_{\sigma_S}(|\overline{p} - \overline{q}|) \cdot G_{\sigma_R}\left(|I_p - I_q|\right) \cdot I_q, \tag{13.1}$$

where $\overline{p}$ and $\overline{q}$ are vectors describing the spatial position of the pixels $\overline{p}$, $\overline{q} \in S$, where S is the spatial domain, the set of all possible pixel positions within the image, I_p and I_q are the intensities of the pixels at positions $\overline{p}$ and $\overline{q}$, I_p, $I_q \in R$, where R is the range or intensity domain, the set of all possible intensities of the image, $G_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ is the Gaussian weighting function, with separate weight parameters σ_S and σ_R for spatial and intensity components, $W_p^b = \sum_{q \in S} G_{\sigma_S}\left(|\overline{p} - \overline{q}|\right) \cdot G_{\sigma_R}\left(|I_p - I_q|\right)$ is the normalization coefficient.

Formula (13.1) states that the resulting intensity I_p^b of the pixel at position $\overline{p}$ is calculated as a weighted sum of intensities of all other pixels in the image with the weights decreasing exponentially with increase of the distance between the pixel at variable position $\overline{q}$ and the reference pixel at position $\overline{p}$. The contributing distances are measured in both spatial and range domains. Owing to the digital nature of the signal, function (13.1) has a finite support and its calculation is truncated to that in the neighborhoods of the pixel at position $\overline{p}$ and intensity I_p. The size of the neighborhood is defined by parameters σ_S and σ_R and sampling rates in both spatial

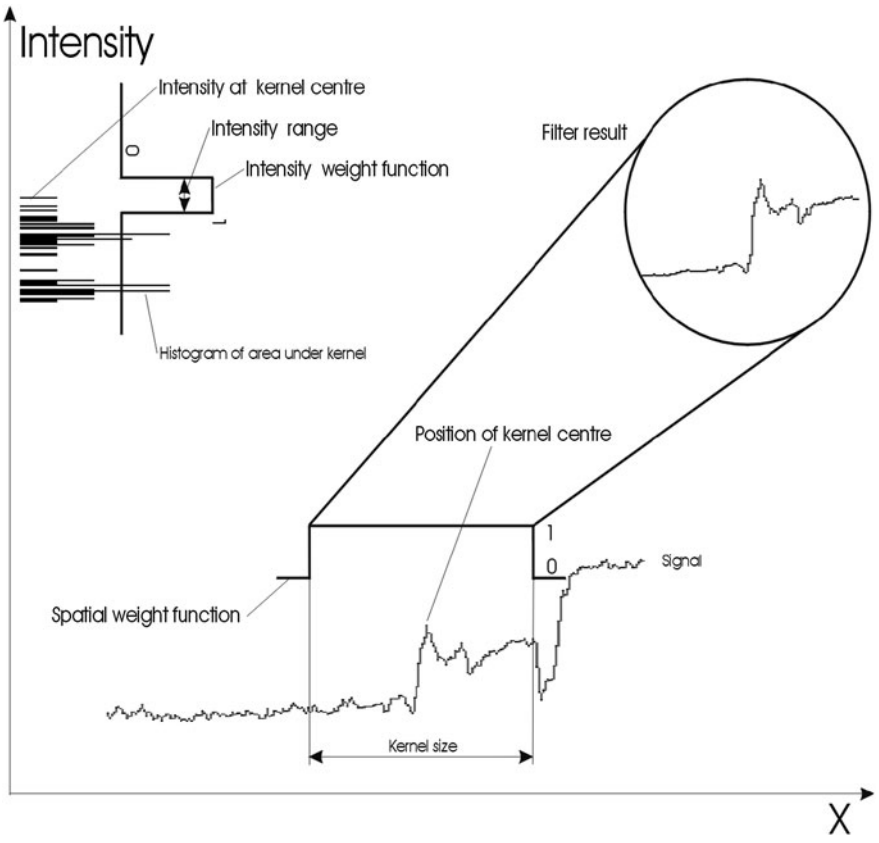


Fig. 13.1 Components making the bilateral filter

and intensity domains. The computation scheme proposed below truncates (13.1) further by giving all pixels in the selected neighborhood the same spatial weight. Furthermore the intensity weighting part of (13.1) applied to the histogram of the neighborhood is reduced to a range constraint around the intensity I_p of the reference pixel. The idea is illustrated by Figs. 13.1 and 13.2. For simplicity a single-dimension signal is presented on the graphs. The components which make the output of the bilateral filter (Fig. 13.1) at a particular spatial position $\overline{p}$ are:

(i) Part of the signal under the kernel centered at the pixel at $\overline{p}$,
(ii) Gaussian spatial weighting function with its maximum at pixel at $\overline{p}$ and "width" parameter σ_S,
(iii) Histogram of the pixels under the kernel centered at pixel at $\overline{p}$,
(iv) Gaussian intensity weighting function with its maximum at I_p and "width" parameter σ_R.

The components which make the proposed filtering scheme (Fig. 13.2) replace the components ii and iv, the Gaussians, with simple windowing functions. The flat

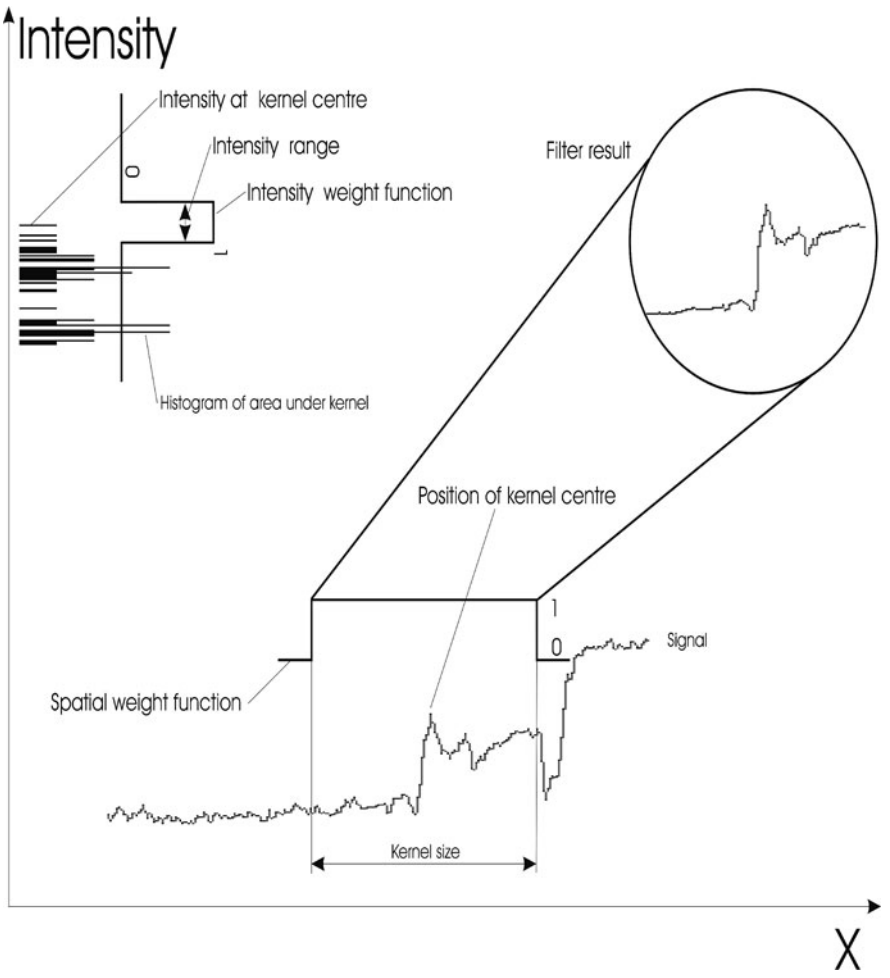


Fig. 13.2 Components making the Intensity constrained flat kernel filtering scheme

kernel works as a spatial filter selecting spatial information in the neighborhood of the reference pixel at $\overline{p}$. This information in the form of a histogram is passed to the intensity filter, which limits the processed information to that in the intensity neighborhood of the reference pixel I_p. This is where the commonality between the bilateral and ICFK filtering schemes ends. For the ICFK scheme the result of the operation depends on the processing function applied to spatially pre-selected data.

$$I_p^{ICFK} = \begin{cases} \mathrm{F}\left(\mathrm{H}_p^{\chi}|_{K(p)}\right), & \text{if} \quad \mathrm{H}_p^{\chi}(I_p) \neq 1 \\ \mathrm{G}\left(\mathrm{H}_p^{\chi}\right), & \text{if} \quad \mathrm{H}_p^{\chi}(I_p) = 1, \end{cases} \tag{13.2}$$

where

H_p^{χ} is a histogram of the part of the image, which is masked by the kernel χ with the centre at $\overline{p}$,

$H_p^{\chi}(I_p)$ is the pixel count of the histogram at the level I_p,

$H_p^{\chi}|_{K(p)}$ is the part of the histogram H_p^{χ} subject to constraint $K(p)$.

Introduction of the second function G, applied only when the intensity level I_p is unique within the region masked by the kernel, is a way of emphasizing the need for special treatment of potential outlayers. Indeed, if the intensity level of a pixel is unique within a sizeable neighborhood, the pixel most likely belongs to noise and should be treated as such.

As will be shown below the selection of functions F and G, as well as the constraint K, defines the nature of the resulting filter, which includes but is not limited by adaptive smoothing.

The output of the filter (13.2) also depends on the shape of the kernel χ. Often in digital image processing, selection of a kernel shape is based on the speed of calculation of filter results as kernel scans across the image. In the case of ICFK filters, this translates into the speed of histogram updates during the scan. There is a significant number of publications [7–9] on methods of speeding up of histogram updates as a square kernel scans the image. In order to avoid shape distortion of the filter output it is more appropriate to use an isotropic kernel, a digital approximation of a circle. A method to speed up the histogram updates while scanning with an isotropic kernel is described in [10]. It is based on the idea proposed in [11]. In the analysis and examples below an isotropic kernel is used. Such a kernel is fully defined by its radius r.

A few words have to be said about the choice of the constraint $K(p)$. In the bilateral filter this role is played by the exponent. By separating the constraint function from the processing functions F and G an extra degree of freedom is added to the filtering scheme. One possible definition of $K(p)$ is offered in Fig. 13.2, where the exponent is replaced by the window function with a fixed window size.

$K(p) = I_p \pm \delta$, where δ is a fixed number that depends on the dynamic range of the source image. For example, for integral image types it is an integer.

In some cases, when looking for dark features on a bright background one may want to employ stronger smoothing to the brighter part of the image and reduce smoothing as the intensity decreases. Then the constraint can take the form

$$K(p) = I_p \pm I_p \cdot \gamma, \tag{13.3}$$

where γ is a fixed ratio.

Furthermore, one can make the constraint adaptive and for example shrink the domain of the function F as the variance within the area masked by the kernel increases:

$$K(p) = I_p \pm [\delta_{\max} - \alpha \cdot (\delta_{\max} - \delta_{\min})],$$

where δ_{max} and δ_{min} are fixed minimum and maximum values for the intensity range,

$$\alpha = \frac{\mathrm{var}(\mathrm{H}_p^{\chi}) - \min_{q \in S}(\mathrm{var}(\mathrm{H}_q^{\chi}))}{\max_{q \in S}(\mathrm{var}(\mathrm{H}_q^{\chi})) - \min_{q \in S}(\mathrm{var}(\mathrm{H}_q^{\chi}))}$$

$$\max_{q \in S}(\mathrm{var}(\mathrm{H}_q^{\chi})) - \min_{q \in S}(\mathrm{var}(\mathrm{H}_q^{\chi})) \neq 0$$

$\mathrm{var}(\mathrm{H}_p^{\chi})$ is the variance of the area under the kernel centered at $\overline{p}$.

13.3 Operators Derived from Intensity Constrained Flat Kernel Filtering Scheme

13.3.1 Edge Preserving Smoothing Filter

This filter can be considered a mapping of the bilateral filter into the ICFK filtering scheme. The functions F and G are given by the following formulae

$$F = \overline{\mathrm{H}_p^{\chi}|_{K(p)}}$$

is the average intensity within that part of the histogram under the kernel mask, which satisfies the constraint $K(p)$,

$$G = median\left(\mathrm{H}_p^{\chi}\right) \tag{13.4}$$

is the median of the area under the kernel mask.

The median acts as a spurious noise suppression filter. From a computational point of view, the update of the histogram as the kernel slides across the image is the slowest operation. It was shown in [10] that the updates of the histogram and the value of the median for an isotropic kernel can be performed efficiently and require $O(r)$ operations, where r is the radius of the kernel.

The edge preserving properties of the filter emanate from the adaptive nature of the function F. The histogram H_p^{χ} is a statistic calculated within the mask of neighborhood χ of the pixel at $\overline{p}$ and comprises intensities of all pixels within that neighborhood. However, the averaging is applied only to the intensities, which are in a smaller intensity neighborhood of I_p constrained by $K(p)$. Thus the output value I_p^{ICFK} is similar in intensity to I_p and intensity-similar features from the spatial neighborhood are preserved in the filter output. If the level I_p is unique in the neighborhood, it is considered as noise and is replaced by the neighborhood median.

An example of the application of the filter is given in Fig. 13.3. The condition (13.3) was used as a constraint. The filter is effective against small particle noise; such as noise produced by camera gain, where linear or median filters would not only blur the edges but would also create perceptually unacceptable noise lumps.

Fig. 13.3 Fragment of an underwater image 733 × 740 pixels with a large number of suspended particles and the result of application of the edge preserving smoothing filter with the radius $r = 12$, subject to intensity constraint $K(p) = Ip \pm Ip \cdot 0.09$

Similarly to the bilateral filter, application of the proposed filter gives the areas with small contrast variation a cartoon-like appearance.

Use of flat kernels for image smoothing was the first choice from the conception of image processing. Other filtering schemes also place some constraints on the pixels within the kernel mask. A good example is the sigma filter [12] and its derivatives [13]. The fundamental difference between the sigma filter and the proposed filter is in the treatment the pixels within the mask. The sigma filter applies the filtering action, mean operator to all the pixels within the mask, if the central pixel is within the certain tolerance, σ range of the mean of the area under the mask, otherwise the filtering action is not applied and the pixel's input value is passed directly to the output. In the proposed filter the Hamlet's question, "to filter or not to filter" is never posed. The filtering action is always applied but only to the pixel subset, which falls within certain intensity range of the central pixel. Moreover the filter output depends only on that, reduced range, not whole region under the kernel mask as in the sigma filter.

13.3.2 Contrast Enhancement Filter for Low Noise Images

The expression (13.2) is general enough to describe not only "smoothing" filters, but "sharpening" ones as well. Consider the following expression for the operator function F:

$$F = \begin{cases} \min\left(\mathrm{H}_p^{\chi}|_{K(p)}\right), & \text{if} \quad I_p \overline{\mathrm{H}_p^{\chi}} \\ \max\left(\mathrm{H}_p^{\chi}|_{K(p)}\right), & \text{if} \quad I_p \geq \overline{\mathrm{H}_p^{\chi}} \end{cases}, \tag{13.5}$$

where $\overline{\mathrm{H}_p^{\chi}}$ is the average intensity of the area under the kernel χ at $\overline{p}$.

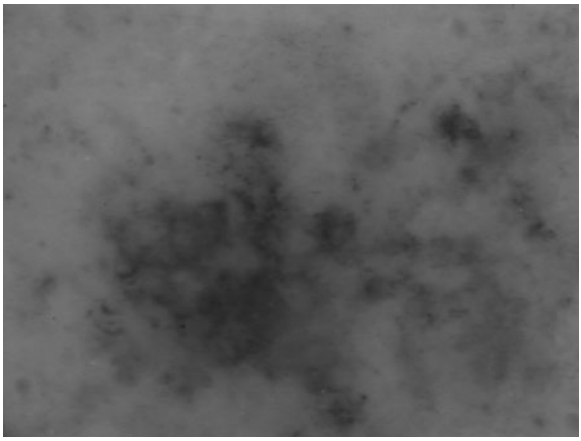


Fig. 13.4 An example of a dermatoscopic image 577 × 434 pixels of a skin lesion

For the purpose of noise suppression the function (13.4) is the recommended choice for G in (13.2). The function F pushes the intensity of the output to one of the boundaries defined by the constraint, depending on the relative position of the reference intensity I_p and the average intensity under the kernel. As any other sharpening operator, the operator (13.5) amplifies the noise in the image. Hence it is most effective on low noise images. Dermatoscopic images of skin lesions can make a good example of this class of images. Dermatoscopy or epiluminescence microscopy is a technique for imaging skin lesions using oil immersion. The latter is employed in order to remove specular light reflection from the skin surface. This technique has a proven diagnostic advantage over clinical photography where images are taken without reflection suppressing oil immersion [14, 15] (Fig. 13.4).

Normally the technique uses controlled lighting conditions. With proper balance of light intensity and camera gain, images taken with digital cameras would have a very low level of electronic noise, while the specular reflection noise is removed by the immersion. An example of such an image is given in Fig. 13.5. Some of the lesions can have a very low inter-feature contrast. Thus both image processing techniques as well as visual inspections can benefit from contrast enhancement. The images in Figs. 13.6 and 13.7 show application of the filter (13.5) and clearly indicate that the constraint parameter γ (13.3) gives a significant level of control over the degree of the enhancement. There is another property of this filter that is worth emphasizing: due to its intrinsic nonlinearity this filter does not produce any ringing at the edges it enhances.

The proposed filter in spirit is not unlike the toggle contrast filter [16]. The difference lies in the degree of contrast enhancement, which in case of the proposed filter has an additional control, the intensity constraint $K(p)$. This control allows making the contrast change as strong as that of toggle contrast filter or as subtle as no contrast change at all.

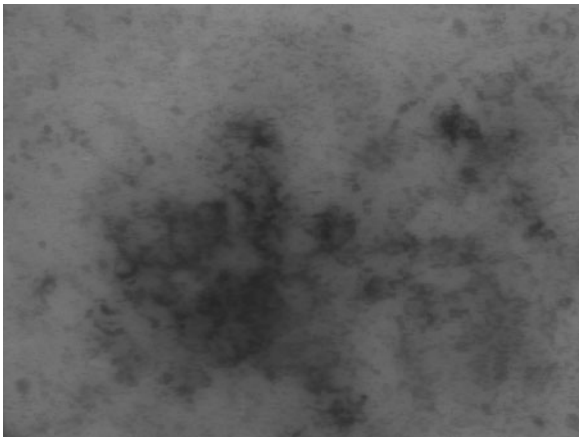


Fig. 13.5 The dermatoscopic image after application of the contrast enhancement filter with the radius $r = 7$, subject to intensity constraint $K(p) = I_p \pm I_p \cdot 0.03$

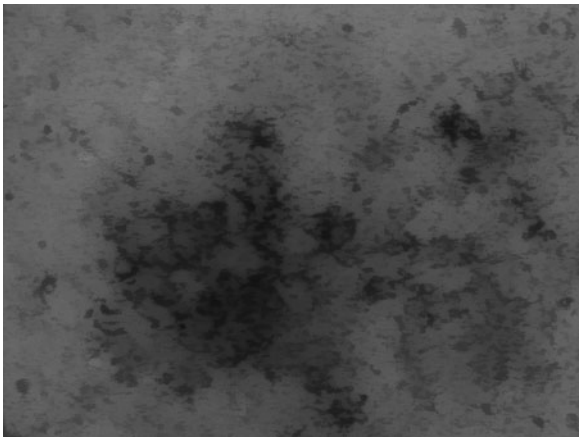


Fig. 13.6 The dermatoscopic image after application of the contrast enhancement filter with the radius $r = 7$, subject to intensity constraint $K(p) = I_p \pm I_p \cdot 0.1$

13.3.3 Local Adaptive Threshold

If sharpening could be considered a dual operation to smoothing and a processing scheme producing a smoothing filter is naturally expected to produce a sharpening one, then here is an example of the versatility of the ICFK scheme and its ability to produce somewhat unexpected operators still falling within the definition (13.2).

Consider a local threshold operator defined by the functions:

$$F = G = \begin{cases} 1, & \text{if} \quad \overline{\mathrm{H}_p^{\chi}} \in \mathrm{H}_p^{\chi}|_{K(p)}, \\ 0, & \text{if} \quad \overline{\mathrm{H}_p^{\chi}} \notin \mathrm{H}_p^{\chi}|_{K(p)} \end{cases} \tag{13.6}$$

where $\overline{\mathrm{H}_p^{\chi}}$ is the average intensity of the area under the kernel χ at $\overline{p}$.

The operator (13.6) produces a binary image, attributing to the background the pixels at which local average for the whole area under the kernel χ at $\overline{p}$ is outside the constrained part of the histogram. The detector (13.6) can be useful in

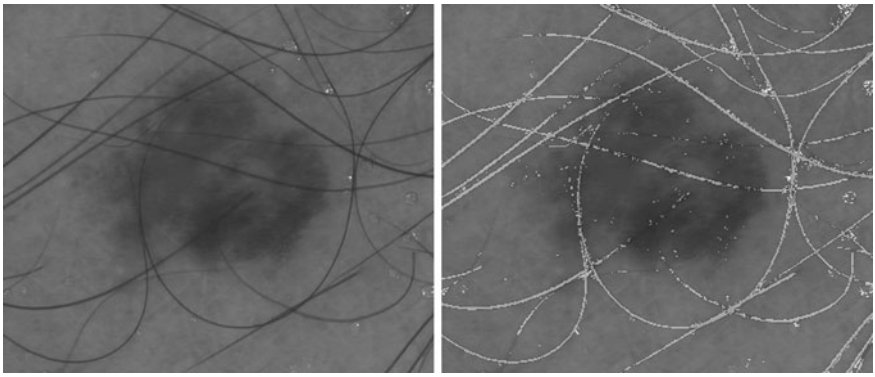


Fig. 13.7 Dermatoscopic image 398 × 339 pixels of a skin lesion with hair and overlay of direct application of the local adaptive threshold with kernel of radius $r = 5$ and intensity constraint $K(p) = I_p \pm I_p \cdot 0.2$

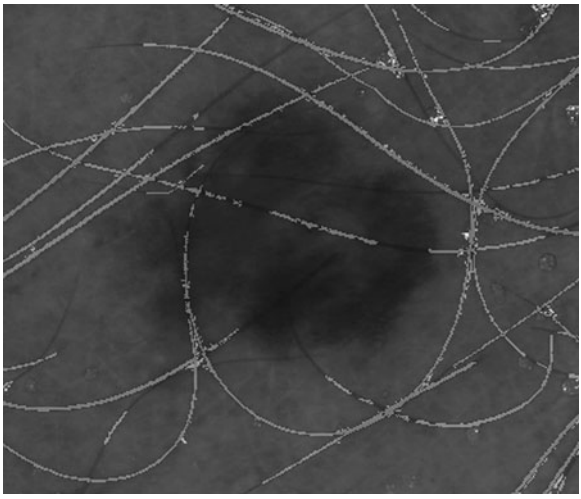


Fig. 13.8 Overlay of application of the local adaptive threshold with kernel of radius $r = 5$ and intensity constraint $K(p) = Ip \pm Ip \cdot 0.2$ followed by morphological cleaning

identifying the narrow linear features in the images. Here is an example, one of the problems in the automatic diagnosis of skin lesions using dermoscopy is removal of artifacts like hairs and oil bubbles trapped in the immersion fluid. The detector (13.6) can identify both of those features as they stand out on the local background. The left half of Fig. 13.7 shows the image with hair and some bubbles. In automated lesion diagnosis systems hair and the bubbles are undesirable artifacts which need to be detected as non-diagnostic features. Prior to application of the operator (13.6) the source image needs to be preprocessed in order to remove the ringing around the hairs caused by sharpening in the video capture device. The preprocessing consists in application of the edge preserving smoothing filter (13.2) with the kernel radius $r = 3$ and the intensity constraint (13.3) where $\gamma = 0.08$. Direct application of filter

(13.6) to the preprocessed image gives the combined hair and bubble mask, which is presented as an overlay on the right of Fig. 13.7. Application of the same filter followed by post-cleaning, which utilizes some morphological operations is presented in Fig 13.8. The advantage of this threshold technique is in its adaptation to the local intensity defined by the size of the processing kernel.

All ICFK filters described above are implemented and available as part of the Pictorial Image Processor© package at www.pic-i-proc.com. The significant part of this work was first presented in [17].

Acknowledgments The author thanks Dr. Scott Menzies from Sydney Melanoma Diagnostic Centre and Michelle Avramidis from the Skintography Clinic for kindly providing dermatoscopic images. Author is also grateful to Prof. H. Talbot for pointing out some similarities between the proposed filters and existing filters.

References

1. Smith SM, Brady JM (1997) SUSAN–a new approach to low level image processing. Int J Comput Vis 23(1):45–78
2. Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: Proceedings of the 1998 IEEE International Conference on Computer Vision. Bombay, India, pp 839–846
3. Kang H, Lee S, Chui CK (2009) Flow based image abstraction. IEEE Trans Vis Comput Graph 16(1):62–76
4. Durand F, Dorsey J (2002) Fast bilateral filtering for the display of high-dynamic-range images. ACM Trans Graph 21(3):257–266
5. Paris S, Durand F (2009) A fast approximation of the bilateral filter using a signal processing approach. Int J Comput Vis 81(1):24–52
6. Elad M (2002) On the bilateral filter and ways to improve it. IEEE Trans Image Process 11(10):1141–1151
7. Gil J, Werman M (1993) Computing 2-D min, median and max. IEEE Trans Pattern Anal Mach Intell 15:504–507
8. Weiss B (2006) Fast median and bilateral filtering. ACM Trans Graph (TOG) 25(3):519–526
9. Perreault S, Hebert P (2007) Median filtering in constant time. IEEE Trans Image Process 16(9):2389–2394
10. Gutenev A, From isotropic filtering to intensity constrained flat kernel filtering scheme. IEEE Trans Image Process (submitted for publication)
11. van Droogenbroeck M, Talbot H (1996) Fast computation of morphological operations with arbitrary structural element. Patt Recog Lett 17:1451–1460
12. Lee JS (1983) Digital image smoothing and the sigma filter. Comp Vis Graph Image Proc 24(2):255–269
13. Lukac R et al (2003) Angular multichannel sigma filter. In: Proceedings. (ICASSP '03) IEEE international conference on acoustics, speech, and signal processing, vol 3, pp 745–748
14. Pehamberger H, Binder M, Steiner A, Wolff K (1993) In vivo epiluminescence microscopy: improvement of early diagnosis of melanoma. J Invest Dermatol 100:356S–362S
15. Menzies SW, Ingvar C, McCarthy WH (1996) A sensitivity and specificity analysis of the surface microscopy features of invasive melanoma. Melanoma Res 6:55–62
16. Kramer HP, Bruckner JB (1975) Iterations of a nonlinear transformation for enhancement of digital images. Pattern Recogn 7:53–58
17. Gutenev AA (2010) Intensity constrained flat kernel image filtering scheme—definition and applications. Lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering, WCE 2010, vol I, 30 June–2 July, London, UK, pp 641–645

Chapter 14
Convolutive Blind Separation of Speech Mixtures Using Auditory-Based Subband Model

Sid-Ahmed Selouani, Yasmina Benabderrahmane, Abderraouf Ben Salem, Habib Hamam and Douglas O'Shaughnessy

Abstract A new blind speech separation (BSS) method of convolutive mixtures is presented. This method uses a sample-by-sample algorithm to perform the subband decomposition by mimicking the processing performed by the human ear. The unknown source signals are separated by maximizing the entropy of a transformed set of signal mixtures through the use of a gradient ascent algorithm. Experimental results show the efficiency of the proposed approach in terms of signal-to-interference ratio (SIR) and perceptual evaluation of speech quality (PESQ) criteria. Compared to the fullband method that uses the Infomax algorithm and to the convolutive fast independent component analysis (C-FICA), our method achieves a better PESQ score and shows an important improvement of SIR for different locations of sensor inputs.

S.-A. Selouani (✉)
Université de Moncton, Shippagan campus, Shippagan, NB E8S 1P6, Canada
e-mail: selouani@umcs.ca

Y. Benabderrahmane · D. O'Shaughnessy
INRS-EMT, Université du Québec, Montreal, H5A 1K6, Canada
e-mail: benab@emt.inrs.ca

D. O'Shaughnessy
e-mail: dougo@emt.inrs.ca

A. B. Salem · H. Hamam
Université de Moncton, Moncton, E1A 3E9, Canada
e-mail: raouf.salem@gmail.com

H. Hamam
e-mail: hamamh@umoncton.ca

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_14,
© Springer Science+Business Media B.V. 2011

14.1 Introduction

The practical goal of blind source separation (BSS) techniques is to extract the original source signals from their mixtures and possibly to estimate the unknown mixing channel using only the information from the observed signal with no, or very limited, knowledge about the source signals and the mixing channel. For several years, the separation of sources has been a particularly active research topic [19]. This interest can be explained by the wide spectrum of possible applications, which includes telecommunications, acoustics, seismology, location and tracking targets of radar and sonar, separation of speakers (so-called 'cocktail party problem'), detection and separation in communication systems for multiple access, etc. Methods to solve the BSS problem can be divided into methods using second-order [5] or higher-order statistics [8], the maximum likelihood principle [3], principle component analysis (PCA) and non-linear PCA [13], and independent component analysis (ICA) methods [11, 14, 18]. Another important category of methods is the subband BSS. Subband BSS has many advantages compared to the other frequency-domain BSS approaches regarding the well-known *permutation ambiguity* of frequency bins [1]. In fact, the subband BSS permutation problem is quite less critical since the number of subbands that could be permuted is obviously smaller than the number of frequency bins. In addition, using a decimation process for each subband can considerably reduce the computational load when compared to time-domain approaches (which could be a computationally demanding task in the case of real-room mixtures). In [2], the subband analysis/synthesis system uses a polyphase filterbank with oversampling and single side band modulation. In low frequency bands, longer unmixing filters with overlap-blockshift are used. In [15], the subband analysis filterbank is basically implemented as a cosine-modulated prototype filter. The latter is designed as a truncated sinc(·) function weighted by a Hamming window. In [20], the impulse responses of the synthesis filters are based on the extended lapped transform and are defined by using the cosine modulation function. In the approach reported in [18], analysis filters are obtained by a generalized discrete Fourier transform. Analysis and synthesis filters are derived from a unique prototype filter which can be designed by iterative least-squares algorithm with a cost function including a stopband attenuation. In the blind speech separation approach we propose, we separate mixed sources that are assumed to be statistically independent, without any a priori knowledge about original source signals $s_j(n),\ j \in \{1, \ldots, N\}$ but using only observations $x_i(n),\ i \in \{1, \ldots, M\}$ through M sensors. Such signals are instantaneously or convolutively mixed. In this work, we are concerned with the convolutive case, i.e. the blind separation of convolved sources of speech, where source signals are filtered by impulse responses $h_{ij}(n)$, from source j to sensor i. We are interested by the indiscriminate approach of separation that offers the advantage of not being reliant on major assumptions on the mix: besides its overall structure, often assumed linear, no settings are supposed known. Mixtures in that case can be expressed under a vector notation as

$$\mathbf{X}(n) = \sum_{k=0}^{\infty} \mathbf{H}(k)\mathbf{S}(n-k), \tag{14.1}$$

where $\mathbf{X}(n) = [x_1(n), \ldots, x_M(n)]^T$ is a vector of mixtures, $\mathbf{S}(n) = [s_1(n), \ldots, s_N(n)]^T$ is a vector of speech sources, and $\mathbf{H}(k) = [h_{ij}(k)], (i,j) \in \{1, \ldots, M\} \times \{1, \ldots, N\}$ is a matrix of FIR filters. To blindly estimate the sources, an unmixing process is carried out, and the estimated sources $\mathbf{Y}(n) = [y_1(n), \ldots, y_N(n)]^T$ can be written as

$$\mathbf{Y}(n) = \sum_{k=0}^{L-1} \mathbf{W}(k)\mathbf{S}(n-k), \tag{14.2}$$

where $\mathbf{W}(k) = [w_{ij}(k)], (i,j) \in \{1, \ldots, M\} \times \{1, \ldots, N\}$ is the unmixing matrix linking the jth output $y_j(n)$ with the ith mixture $x_i(n)$. Such matrix is composed of FIR filters of length L. Each element is defined by the vectors $w_{ij}(k) = [w_{ij}(0), \ldots, w_{ij}(L-1)], \forall (i,j) \in \{1, \ldots, M\} \times \{1, \ldots, N\}$.

To mitigate problems in both time and frequency domains, in next sections, a new framework for the BSS of convolutive mixtures based on subband decomposition using an ear-model based filterbank and information maximization algorithm is presented and evaluated. This chapter includes an extension of our previous work [6]. A new evaluation criteria is introduced, namely the PESQ, and a new set of experiments are carried out involving the well-known C-FICA method in different mixing conditions.

14.2 Proposed Method

In this section, we introduce the convolutive mixture based on the head related transfer function (HRTF) that we used to evaluate the proposed method. Then we define the subband decomposition using the modeling of the mid-external ear and the basilar membrane that aims at mimicking the human auditory system (HAS). Afterwards, the learning rule performing the sources' separation is introduced.

14.2.1 HRTF Mixing Model

The perception of the acoustic environment or room effect is a complex phenomenon linked mainly to the multiple reflections, attenuation, diffraction and scattering on the constituent elements of the physical environment around the sound source that the acoustic wave undergoes in its propagation from source to ear. These phenomena can be modeled by filters representing diffraction, scattering and reflection that a sound wave sustains during its travel between its source and the entrance of the ear canal of the listener. These filters are commonly called the head related transfer function or HRTF [10]. The principle of measuring HRTF is to place microphones in the ears and record the signals corresponding to different source positions. The HRTF is the

transfer function between the source signals and the signals at the ears. The HRTF is then considered as a linear and time-invariant system. Each HRTF is represented by an FIR filter (finite impulse response), causal and stable. In our experiments, sources are convoluted with impulse responses modeling the HRTF. We tested our overall framework with mixing filters measured at the ears of a dummy head. We selected impulse responses associated with source positions defined by various angle values in relation to the dummy head (see Fig. 14.4).

14.2.2 Subband Decomposition

The proposed modeling of HAS consists of three parts that simulate the behavior of the mid-external ear, the inner ear and the hair-cells and fibers. The external and middle ear are modeled using a bandpass filter that can be adjusted to signal energy to take into account the various adaptive motions of ossicles. The model of inner ear simulates the behavior of the basilar membrane (BM) that acts substantially as a non-linear filter bank. Due to the variability of its stiffness, different places along the BM are sensitive to sounds with different spectral content. In particular, the BM is stiff and thin at the base, but less rigid and more sensitive to low frequency signals at the apex. Each location along the BM has a characteristic frequency, at which it vibrates maximally for a given input sound. This behavior is simulated in the model by a cascade filter bank. The number of filterbank depends on the sampling rate of the signals and on other parameters of the model such as the overlapping factor of the bands of the filters, or the quality factor of the resonant part of the filters. The final part of the model deals with the electro-mechanical transduction of hair-cells and afferent fibers and the encoding at the level of the synaptic endings [7, 21].

14.2.2.1 Mid-External Ear

The mid-external ear is modeled using a bandpass filter. For a mixture input $x_i(k)$, the recurrent formula of this filter is given by

$$x'_i(k) = x_i(k) - x_i(k-1) + \alpha_1 x'_i(k-1) - \alpha_2 x'_i(k-2), \tag{14.3}$$

where $x'_i(k)$ is the filtered output, $k = 1, \ldots, K$ is the time index and K is the number of samples in a given block. The coefficients α_1 and α_2 depend on the sampling frequency F_s, the central frequency of the filter and its Q-factor.

14.2.2.2 Mathematical Model of the Basilar Membrane

After each frame is transformed by the mid-external filter, it is passed to the cochlear filter banks whose frequency responses simulate those of the BM for an auditory stimulus in the outer ear. The formula of the model is as follows:

$$x_i''(k) = \beta_{1,i} x_i''(k-1) - \beta_{2,i} x_i''(k-2) + G_i[x_i'(k) - x_i'(k-2)], \tag{14.4}$$

and its transfer function can be written as:

$$H_i(z) = \frac{G_i(1 - z^{-2})}{1 - \beta_{1,i} z^{-1} + \beta_{2,i} z^{-2}}, \tag{14.5}$$

where $x_i''(k)$ is the BM displacement which represents the vibration magnitude at position δ_i and constitutes the BM response to a mid-external sound stimulus $x_i'(k)$. The parameters G_i, $\beta_{1,i}$ and $\beta_{2,i}$, respectively the gain and coefficients of filter or channel i, are functions of the position δ_i along the BM. N_c cochlear filters are used to realize the model. These filters are characterized by the overlapping of their bands and a large bandwidth. The BM has a length of 35 mm which is approximately the case for humans [7]. Thus, each channel represents the state of an approximately $\Delta = 1.46$ mm of the BM. The sample-by-sample algorithm providing the outputs of the BM filters is given as follows.

Sample-by-sample algorithm

Initialize $f_x = (F_s\, \Delta x)^2; H_0 = 0; r_{i,j} = 0; E_0 = 0.$
For $i = 1$ **to** N_c **do**
$x_i = i\, \Delta x; v = e^{-106.5\, x_i}; F_i = 7100\, v - 100;$
$C_i = \frac{(27\, v)^2}{f_x}; Q_i = (-8300\, x_i + 176.3)\, x_i + 4;$
$G_i = e^{-80\, x_i}; u = e^{-\frac{\pi\, F_i}{F_s\, Q_i}}; \beta_{1,i} = 2\, u\, cos(\frac{2\, \pi\, F_i}{F_s});$
$\beta_{2,i} = u^2; E_i = \frac{1}{1+(2-E_{i-1})\, C_i}; A_i = E_i\, C_i.$
EndDo
For $k = 1$ **to** K **Do**
For $i = 1$ **to** N_c **Do**
$H_i = [G_i\, (s'(k) - s'(k-2)) + \beta_{i,2}\, r_{1,i} -$
$\beta_{2,i}\, r_{i,1}]\, E_i + H_{i-1}\, A_i$
EndDo
For $i = 1$ **to** N_c **Do**
$r_{1,i} = A_i\, r_{i+1,3} + H_i; y'_i(k) = r_{i,3}$
EndDo
For $i = 1$ **to** N_c **Do**
For $j = 1$ **to** 2 **Do**
$r_{i,j} = r_{i,j+1}$
EndDo
EndDo
EndDo

14.2.3 Learning Algorithm

After performing the subband decomposition, the separation of convolved sources per subband is done by the Infomax algorithm. Infomax was developed by Bell and Sejnowski for the separation of instantaneous mixtures [4]. Its principle consists of maximizing output entropy or minimizing the mutual information between components of $\mathbf{Y}$ [23]. It is implemented by maximizing, with respect to $\mathbf{W}$, the entropy of $\mathbf{Z} = \Phi(\mathbf{Y}) = \Phi(\mathbf{WX})$. Thus, the Infomax contrast function is defined as

$$C(\mathbf{W}) = H(\Phi(\mathbf{WX})), \tag{14.6}$$

where $H(\cdot)$ is the differential entropy, which can be expressed as $H(a) = -E[Ln(f_a(a))]$, where $f_a(a)$ denotes the probability density function of a variable a. The generalization of Infomax for the convolutive case is performed by using a feedforward architecture. Both causal and non-causal FIR filters are performed in our experiments. With real-valued data for vector $\mathbf{X}$, entropy maximization algorithm leads to the adaptation of unmixing filter coefficients with a stochastic gradient ascent rule using a learning steepest μ. Then, the weights are defined as follows:

$$\mathbf{W}(0) = \mathbf{W}(0) + \mu([\mathbf{W}(0)]^{-T} - \Phi(\mathbf{Y}(n))\mathbf{X}^T(n)), \tag{14.7}$$

and,

$$w_{ij}(k) = w_{ij}(k) - \mu\Phi(y_i(n))x_j(n-k); \quad \forall k \neq 0, \tag{14.8}$$

where $\mathbf{W}(0)$ is a matrix composed of unmixing FIR filters coefficients as defined in Sect. 14.1, $\mathbf{Y}(n)$ and $\mathbf{X}(n)$ are the separated sources and the observed mixtures, respectively. $\Phi(\cdot)$ is the score function of y_i which is a non-linear function approximating the cumulative density function of sources, as defined in Eq. 14.9, where $p(y_i)$ denotes the probability density function of y_i

$$\Phi(y_i(n)) = \frac{\frac{\delta p(y_i(n))}{\delta y_i(n)}}{p(y_i(n))}. \tag{14.9}$$

The block diagram of the proposed method is given in Fig. 14.1. The input signals, that are the set of mixtures, are firstly processed by the mid-external ear introduced by Eq. 14.3. Then outputs are passed through a filterbank representing the cochlear part of the ear. A decimation process is then performed for each subband output. Such decimation is useful for many reasons. First, it improves the convergence speed because input signals are more whitened than the time domain approach. Second, the wanted unmixing filter length will be reduced by a factor of $\frac{1}{M}$, where M is the decimation factor. After performing decimation, we group a set of mixtures belonging to the same cochlear filter to be the input of the unmixing stage. The latter gives separated sources of each subband that are upsampled by a

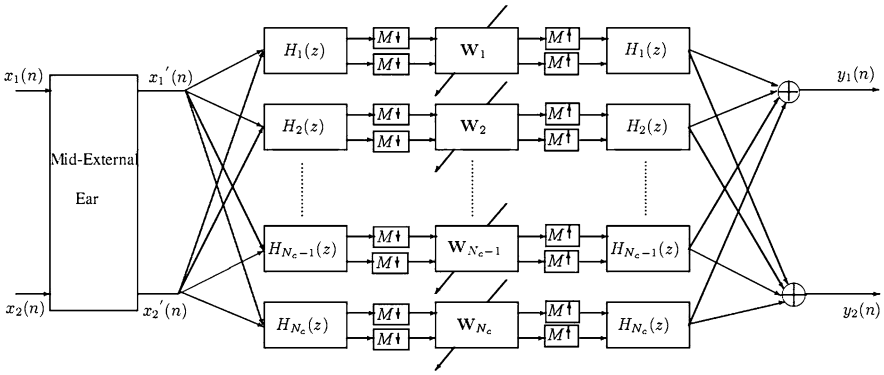


Fig. 14.1 The ear-based framework for the subband BSS of convolutive mixtures of speech

M factor. The same filter bank is used for the synthesis stage. The estimated sources are added from different synthesis stages.

14.3 Experiments and Results

A set of nine different signals, consisting of speakers (three females and six males) reading sentences during approximately 30 s, was used throughout experiments. This speech signals were collected by Nion et al. [17]. The signals were down-sampled to 8 kHz. The C-FICA algorithm (convolutive extension of Fast-ICA: independent component analysis) and the full-band Infomax algorithms are used as baseline systems for evaluation. The C-FICA algorithm proposed by Thomas et al. [22] consists of time-domain extensions of the fast-ICA algorithms developed by Hyvarinen et al. [11] for instantaneous mixtures. For an evaluation of the source contributions, C-FICA uses the criterion of least squares, whose optimization is carried out by a Wiener filtering process. The convolutive version of full-band Infomax introduced in Sect. 14.3 in the evaluation tests.

14.3.1 Evaluation Criteria

To evaluate the performance of BSS methods, two objective measures were used namely the signal to interference ratio (SIR) and the perceptual evaluation of speech quality (PESQ). The SIR has been emphasized to be a most efficient criterion for several methods aiming at reducing the effects of interference [9]. The SIR is an important entity in communications engineering that indicates the quality of a speech signal between a transmitter and a receiver environment. It is selected as the criteria for optimization. This reliable measurement is defined by

$$\mathrm{SIR} = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2}, \tag{14.10}$$

where $s_{target}(n)$ is an allowed deformation of the target source $s_i(n)$, $e_{interf}(n)$ is an allowed deformation of the sources which accounts for the interference of the unwanted sources. Those signals are derived from a decomposition of a given estimated source $y_i(n)$ of a source $s_i(n)$.

The second measure used to evaluate the quality of source separation is the PESQ. The latter is normalized in ITU-T recommendation P.862 [12] and is generally used to evaluate speech enhancement systems [16]. Theoretically, the results can be mapped to relevant mean opinion scores (MOS) based on the degradation of the speech sample. The algorithm predicts subjective opinion scores for degraded speech samples. PESQ returns a score from 0.5 to 4.5. The higher scores suggest better quality. The code provided by Loizou in [16] is used in our experiments. In general, the reference signal indicates an original signal and the degraded signal indicates the same utterance pronounced by the same speaker as in the clean signal but submitted to diverse adverse conditions. In the PESQ algorithm, the reference and degraded signals are level-equalized to a standard listening level thanks to the pre-processing stage. The gain of the two signals may vary considerably, so it is a priori unknown. In the original PESQ algorithm, the gains of the reference, degraded and corrected signals are computed based on the root mean square values of band-passed-filtered (350–3,250 Hz) speech.

14.3.2 Discussion

Different configurations of the subband analysis and synthesis stages as well as of the decimation factor have been tested. The number of subbands was fixed at 24. Through our experiments we observed that when we keep the whole number of subbands, the results were not satisfactory. In fact, we noticed that some subbands in high frequencies are not used, and therefore this causes distortions on the listened signals. However, as shown in Fig. 14.2, the best performance was achieved for $N_c' = 24$ and $M = 4$. In addition to the use of causal FIR filters, we adapted unmixing stage weights for non-causal FIR by centering the L taps. From Fig. 14.2, we observe that causal FIR yields good results in SIR improvement when compared to non-causal one. Another set of experiments have been carried out to evaluate the performance in the presence of an additive noise in sensors. We used the signal-to-noise-ratio (SNR) which is defined in [9], by

$$\mathrm{SNR} = 10\log_{10}\frac{||s_{target} + e_{interf}||^2}{||e_{noise}||^2}, \tag{14.11}$$

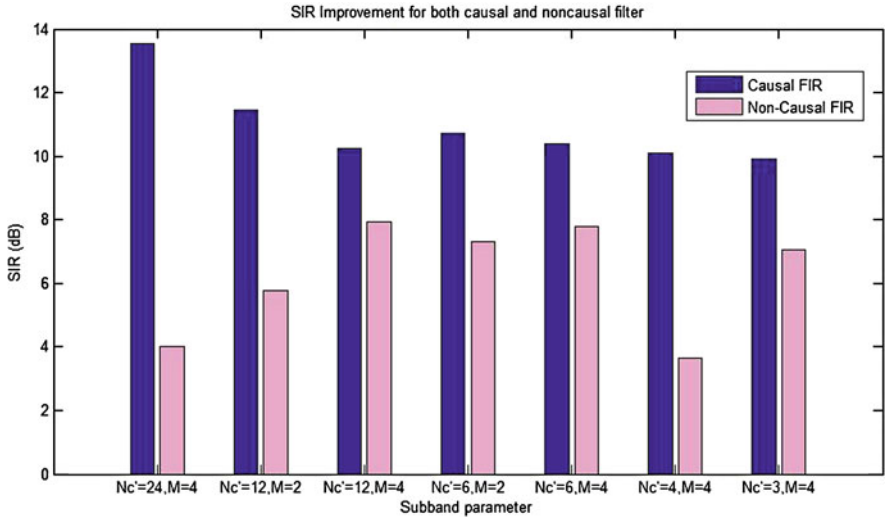


Fig. 14.2 SIR improvement for both causal and noncausal filters. We denote by N_c' the number of filters that have been used among N_c filters and M is decimation factor

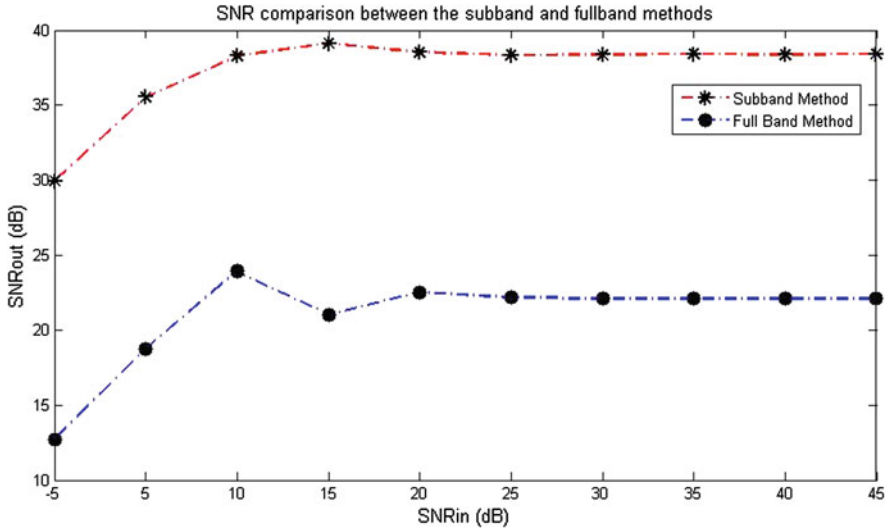


Fig. 14.3 SNR comparison between the subband and fullband methods

where e_{noise} is an allowed deformation of the perturbating noise, s_{target} and e_{interf} were defined previously. Figure 14.3 shows the SNR improvement using our subband decomposition, comparing to the fullband method, i.e. Infomax algorithm in convolutive case.

We have also compared the proposed method with the well-known C-FICA and fullband Infomax techniques using PESQ and SIR objective measures. Among the available data, we considered a two-input, two-output convolutive BSS problem. We mixed in convolution two speech signals pronounced by a man and a woman. We repeated this procedure with different couples of sentences. The average of evaluation measures (SIR and PESQ) were calculated. As illustrated in Fig. 14.4, we selected impulse responses associated with source positions defined by different angles in relation to the dummy head. As can be seen in Table 14.1, the proposed subband is efficient and has the additional advantage that the preprocessing step is not necessary. The method was also verified subjectively by listening to the original, mixed and separated signals. The PESQ scores confirm the superiority of the proposed method in terms of intelligibility and quality of separation when compared to baseline techniques. The best HRTF configuration

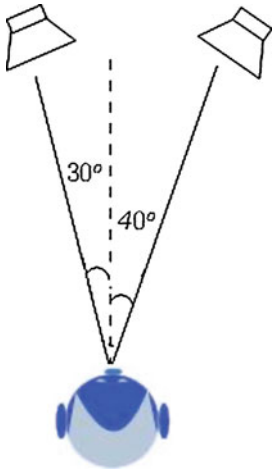


Fig. 14.4 The convolutive model with source positions at 30 and 40° angles in relation to the dummy head

Table 14.1 SIR and PESQ of proposed subband BSS, C-FICA method and full-band BSS

Angle (°)	C-FICA		Full-band BSS		Proposed	
	PESQ	SIR	PESQ	SIR	PESQ	SIR
10	2.61	5.08	3.15	8.45	3.85	13.45
10	2.68	5.37	3.27	9.74	4.10	13.62
10	2.04	4.45	2.23	7.21	2.92	8.96
60	2.52	6.67	2.76	8.94	3.24	10.05
20	3.18	7.54	3.81	11.28	4.24	13.83
50	2.95	6.82	3.55	10.79	4.16	11.67
20	3.02	6.62	3.32	10.02	3.42	12.14
120	2.15	5.87	3.04	9.11	3.29	10.52
30	2.28	6.29	2.47	7.55	2.88	9.76
80	1.94	4.72	2.13	7.02	2.51	8.73

was obtained for 20–50° angle of dummy head where a PESQ of 4.24 and a SIR of 13.83 dB were achieved.

14.4 Conclusion

An ear-based subband BSS approach was proposed for the separation of convolutive mixtures of speech. The results showed that using a subband decomposition that mimics the human perception and using the Infomax algorithm yields better results than the fullband and C-FICA methods. Experimental results showed the high efficiency of the new method in improving the SNR of unmixed signals in the case of noisy sensors. It is worth noting that an important advantage of the proposed technique is that it uses a simple time-domain sample-by-sample algorithm to perform the decomposition and that it does not need pre-processing step.

References

1. Araki S, Makino S, Nishikawa T, Saruwarati H (2001) Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech. In: IEEE-ICASSP conference, pp 2737–2740
2. Araki S, Makino S, Aichner R, Nishikawa T, Saruwatari H (2005) Subband-based blind separation for convolutive mixtures of speech. IEICE Trans Fundamentals E88-A(12):3593–3603
3. Basak J, Amari S (1999) Blind separation of uniformly distributed signals: a general approach. IEEE Trans Neural Networks 10:1173–1185
4. Bell AJ, Sejnowski TJ (1995) An information maximization approach to blind separation and blind deconvolution. Neural Comput 7(6):1129–1159
5. Belouchrani A, Abed-Meraim K, Cardoso JF, Moulines E (1997) A blind source separation technique using second-order statistics. IEEE Trans Signal Process 45(2):434–444
6. Ben Salem A, Selouani SA, Hamam H (2010) Auditory-based subband blind source separation using sample-by-sample and Infomax algorithms. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering, 2010, WCE 2010, 30 June–2 July, London, UK, pp 651–655
7. Caelen J (1985) Space/time data-information in the A.R.I.A.L. project ear model. Speech Commun J 4:163–179
8. Cardoso JF (1989) Source separation using higher order moments. In: Proceedings IEEE ICASSP, Glasgow, UK, vol 4, pp 2109–2112
9. Fevotte C, Gribonval R, Vincent E (2005) BSS_EVAL toolbox user guide. IRISA, Rennes, France, Technical Report 1706 [Online]. Available: http://www.irisa.fr/metiss/bss_eval
10. Gardner B, Martin K Head related transfer functions of a dummy head [Online]. Available: http://www.sound.media.mit.edu/ica-bench/
11. Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York
12. ITU (2000) Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation 862
13. Karhunen J, Joutsensalo J (1994) Representation and separation of signals using nonlinear PCA type learning. Neural Networks 7:113–127

14. Kawaguchi A (2010) Statistical inference for independent component analysis based on polynomial spline model. In: IANG conference, vol I, IMECS 2010, Hong Kong
15. Kokkinakis K, Loizou PC (2007) Subband-based blind signal processing for source separation in convolutive mixtures of speech. In: IEEE-ICASSP conference, pp 917–920
16. Loizou P (2007) Speech enhancement: theory and practice. CRC Press LLC, Boca Raton
17. Nion D, Mokios KN, Sidiropoulos ND, Potamianos A (2010) Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures. IEEE Trans Audio Speech Language Process 18(6):1193–1207
18. Park HM, Dhir CS, Oh SH, Lee SY (2006) A filter bank approach to independant component analysis for convolved mixtures. Neurocomputing 69:2065–2077
19. Pedersen MS, Larsen J, Kjems U, Parra LC (2007) A survey of convolutive blind source separation methods. Springer handbook on speech processing and speech communication. Springer, Berlin
20. Russel I, Xi J, Mertins A, Chicharo J (2004) Blind source separation of non-stationary convolutively mixed signals in the subband domain. In: IEEE-ICASSP conference, pp 481–484
21. Tolba H, Selouani SA, O'Shaughnessy D (2002) Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multistream paradigm. In: IEEE-ICASSP conference 2002, pp 837–840
22. Thomas J, Deville Y, Hosseini S (2006) Time-domain fast fixed-point algorithms for convolutive ICA. IEEE Signal Process Lett 13(4):228–231
23. Wax M, Kailath T (1985) Detection of signals by information theoretic criteria. IEEE Trans Acoust Speech Signal Process 33(2):387–392

Chapter 15
Time Domain Features of Heart Sounds for Determining Mechanical Valve Thrombosis

Sabri Altunkaya, Sadık Kara, Niyazi Görmüş and Saadetdin Herdem

Abstract Thrombosis of implanted heart valve is a rare but lethal complication for patients with mechanical heart valve. Echocardiogram of mechanical heart valves is necessary to diagnose valve thrombosis definitely. Because of the difficulty in making early diagnosis of thrombosis, and the cost of diagnosis equipment and operators, improving noninvasive, cheap and simple methods to evaluate the functionality of mechanical heart valves are quite significant especially for first step medical center. Because of this, time domain features obtained from auscultation of heart sounds are proposed to evaluate mechanical heart valve thrombosis as a simple method in this chapter. For this aim, heart sounds of one patient with mechanical heart valve thrombosis and five patients with normally functioning mechanical heart valve were recorded. Time domain features of recorded heart sounds, the skewness and kurtosis, were calculated and statistically evaluated using paired and unpaired *t*-test. As a result, it is clearly seen that the skewness of first heart sound is the most discriminative features ($p < 0.01$) and it

S. Altunkaya (✉) · S. Herdem
Department of Electrical and Electronics Engineering, Selçuk University,
42075 Konya, Turkey
e-mail: saltunkaya@selcuk.edu.tr

S. Herdem
e-mail: sherdem@selcuk.edu.tr

S. Kara
Biomedical Engineering Institute, Fatih University, 34500 Istanbul, Turkey
e-mail: skara@fatih.edu.tr

N. Görmüş
Department of Cardiovascular Surgery, Meram Medical School of Selcuk University,
42080 Konya, Turkey
e-mail: ngormus@selcuk.edu.tr

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*,
Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_15,
© Springer Science+Business Media B.V. 2011

may be used fairly well in differentiating normally functioning mechanical heart valve from malfunctioning mechanical heart valve.

15.1 Introduction

Mechanical heart valve thrombosis is any thrombosis attached to a mechanical valve, occluding part of the blood flow or interfering with valvular function [1]. The mechanical heart valve thrombosis is a critical complication relating to the high mortality and requires immediate diagnosis and thrombolytic or surgical treatment [2]. Progression in the structure and design of mechanical heart valve over the years has led to a considerable improvement in their hemodynamic features and durability. However, thromboembolic complications remain a troublesome cause of postoperative morbidity and mortality [3, 4]. According to different literature; incidence of thromboembolic complication ranges from 0.03 to 4.3% patient-years [2], 0.5–6% per patient-year [3], 2–4% patients per annum [5] 0.5% patient-years [6] depending on the generation and the thrombosis of the prosthesis used, the location of the valve, and the quality of the anticoagulation [2]. Recently, transesophageal echocardiography has become the gold standard both in the early diagnosis of prosthetic valve thrombosis and in risk stratification for obstruction or embolism in patients with prosthetic heart valves [7]. However, it is quite expensive to use echocardiography for diagnosis of mechanical heart valve thrombosis in the first step medical center because of both a specialist requirement to use echocardiography and cost of these devices. So it may cause misdiagnosis of thrombotic complications in the first step medical center. Therefore, improving noninvasive, cheap and simple methods to evaluate the functionality of the mechanical heart valve are quite important [8–11].

There are limited numbers of study to evaluate mechanical heart valve thrombosis using heart sounds. Although there are limited numbers of studies about frequency spectrum of mechanical heart valve sounds, it is known that thrombosis formation on a prosthetic heart valve changes the frequency spectrum of both biological and mechanical heart valve. Features obtained from frequency and time–frequency analysis of heart sounds are used to detect mechanical heart valve thrombosis in the past studies. In these past studies, generally modified forward–backward Proony's Method was used to detect frequency component of prosthetic heart valve [8–10, 12, 13].

In this chapter, time domain features instead of frequency domain features are proposed to evaluate thrombosis on the mechanical heart valve. For this aim, heart sounds of one patient with thrombosis and heart sounds of five patients with normally functioning mechanical heart valve are recorded. The skewness and kurtosis of heart sounds as time domain features were found and statistically evaluated using *t*-test.

Table 15.1 Clinical information of patients

Pat. no	Sex	Age	Valve size (mm)	Valve type	Condition
1	F	58	25	Sorin	Normal
2	F	27	29	Sorin	Normal
3	F	35	29	Sorin	Normal
4	F	30	29	St.Jude	Normal
5	F	45	29	St.Jude	Normal
6	F	55	29	Sorin	With thrombosis

15.2 Patients and Data Acquisition

This study includes patients who were operated in the Department of Cardiovascular Surgery, Meram Medical School of Selcuk University. Five patients with normally functioning mechanical heart valve and one patient with mechanical valve thrombosis were selected to evaluate the mechanical heart valve thrombosis using heart sounds. The heart sounds of a patient with thrombosis were recorded before and after thrombolytic treatment. Functionality of the mechanical heart valve of patients was investigated using echocardiography by the physician. After echocardiography investigation, thrombus with partial obstruction was monitored on the mitral mechanical heart valve of a patient. The heart sounds of five patients with normally functioning mechanical heart valve were recorded after the heart valve replacement in one year. The heart sounds of patients recorded from mitral area (intersection of left 5. intercostals interval and mid clavicular line) over the entire course of 30 s [14]. All patients had mitral valve replacement and clinical information of these patients is shown on Table 15.1.

ECG signals were recorded simultaneously with heart sounds to segment first and second heart sounds. E-Scope II electronic stethoscope manufactured by Cardionics was used to record heart sounds. Sound signals obtained from electronic stethoscope and ECG signals obtained from the surface electrode were digitized at a 5000 Hz sampling frequency using the Biopac MP35 data acquisition device.

15.3 Extraction of First Heart Sounds (S1) and Second Heart Sounds (S2)

As mentioned in the previous section, 30 s heart sounds are recorded from each patient. One S1 and one S2 sound component available in the heart sounds signal for one heart beat. In this chapter, detection of S1 and S2 that is varying number according to the number of pulse are discussed. Known that, the S1 occurs after the onset of the QRS complex, the S2 occurs towards the end of the T wave of ECG. Using these two relations between heart sounds and ECG, S1 and S2 sounds obtained from 30 s record. Processing of recorded heart sounds signal can be

summarized as follows. Firstly, filtration and normalization of recorded heart sounds and ECG signal is performed. After that, QRS and T peak of ECG signal is detected. Finally, S1 and S2 sounds are detected using QRS and T peak [14].

15.3.1 Preprocessing of ECG and Heart Sounds Signals

All recorded heart sounds were filtered with a 30 Hz high pass and 2000 Hz low pass digital finite impulse response filter to get rid of noise and were normalized using

$$HS_{norm}(n) = \frac{HS(n)}{max|HS(n)|} \tag{15.1}$$

where $HS(n)$ is the row heart sound signal and $HS_{norm}(n)$ is the normalized heart sounds signal. Also, a normalizing process was applied to the ECG signal.

15.3.2 Detection of QRS Complex and T Peak

QRS complex of the ECG signal are detected using a first-derivative based QRS detection algorithm. In this algorithm, the ECG signal is first band pass filtered with a pass band of 10–20 Hz to eliminate the baseline wander and high frequency noise. After filtering, the ECG signal is differentiated to obtain QRS complex slope information, is squared point by point to clarify the QRS complex in the ECG signal, and then is time-averaged by taking the mean of the previous 10 points. The time-averaged ECG signal is compared to a threshold to obtain the QRS complex [15, 16]. The threshold is chosen to be a quarter of the maximum time averaged ECG signal. As a result of this comparison, the maximum of the time-averaged ECG signals greater than the threshold is accepted as R peak. After that, all intervals between consecutive R peaks (RR interval) are compared to 0.5 and 1.5 times the mean RR interval. If the RR interval is longer than 1.5 times of the mean RR interval and is shorter than 0.5 times the mean RR interval, then this RR interval and its counterpart of heart sounds is removed from a signal to prevent wrong detection of RR interval.

T waves are detected using the physiological knowledge that the peak in the T wave occurs at least 60 ms after the R peak and is normally within the two-thirds of the RR interval. The maximum of ECG signal in these interval is used as a T peak to detect the location of S2 [17].

15.3.3 Detection of S1 and S2

The ECG signal and Shannon energy is used to detection of the heart sounds. The ECG signal is used as a time reference to determine the time interval

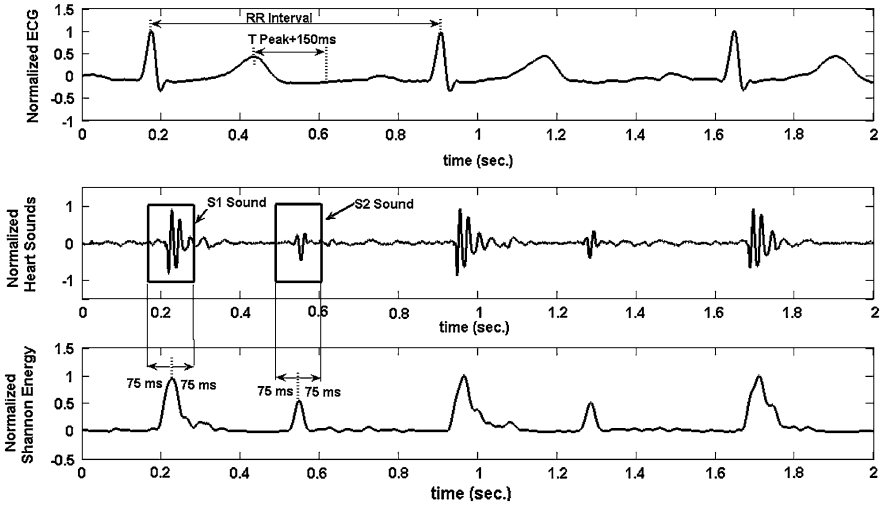


Fig. 15.1 ECG, heart sounds and Shannon energy of one patient with MVR

where S1 and S2 are searched over one heart cycle. The Shannon energy of heart sounds is used for exact determination of location of S1 and S2 in the finite interval. The Shannon energy of the normalized heart valve sound (HS_{norm}) can be calculated using

$$SE = -\frac{1}{N}\sum_{n=1}^{N} HS_{norm}^2(n) \cdot \log HS_{norm}^2(n) \tag{15.2}$$

where SE is Shanonn energy of HS_{norm}, N is length of recorded data and n is index of HS_{norm} [18].

After the Shannon energy of heart sounds is calculated, to determine exactly the location of S1 in the RR interval, the maximum point of the Shannon energy in the interval between 0 01 RR and 0.2 RR is accepted the center of the S1. The maximum point of Shannon's energy between the times the ECG T peaks to the ECG T peak time plus 150 ms is accepted as the center of S2. The duration of S1 and S2 was chosen to be 150 and 75 ms respectively on both sides of the center (Fig. 15.1). If the Shannon energy of the right or left side of the center is larger than 40% of the maximum Shannon energy, the duration of chosen heart sounds is increased by 20%. The comparison is repeated until the Shannon energy of the right or left side of the center is smaller than 40% the maximum Shannon energy [17, 19]. In Fig. 15.1, the upper graph shows the RR interval of the ECG signal, the 0.3–0.65 RR interval, and between the times the ECG T peaks to the ECG T peak time plus 150 ms, the middle graph shows the heart sounds signal, and the bottom graph shows the calculated Shannon energy.

Table 15.2 Mean ± Standard deviation (std.) of the skewness and kurtosis of heart sounds

	Thr (mean ± std.)	AThr (mean ± std.)	N (mean ± std.)
Skewness of S1	0.96±0.36	0.18±0.25	0.12±0.42
Skewness of S2	0.71±0.7	0.3±0.36	−0.2±0.54
Kurtosis of S1	5.24±1.11	4.34±0.45	5.9±1.75
Kurtosis of S2	8.39±2.47	5.65±1.56	5.79±1.92

15.4 Skewness and Kurtosis

The change in the signal or distribution of the signal segments is measured in terms of the skewness and kurtosis. The skewness characterizes the degree of asymmetry of a distribution around its mean. The skewness is defined for a real signal as

$$Skewness = \frac{E(x-\mu)^3}{\sigma^3} \tag{15.3}$$

where μ are the mean and σ are the standard deviation and E denoting statistical expectation. The skewness shows that the data are unsymmetrically distributed around a mean. If the distribution is more to the right of the mean point the skewness is negative. If the distribution is more to the left of the mean point the skewness is positive. The skewness is zero for a symmetric distribution. The kurtosis measures the relative peakedness or flatness of a distribution. The kurtosis for a real signal $x(n)$ is calculated using

$$Kurtosis = \frac{E(x-\mu)^4}{\sigma^4} \tag{15.4}$$

where μ are the mean and σ are the standard deviation and E denotes statistical expectation. For symmetric unimodal distributions, the kurtosis is higher than 3 indicates heavy tails and peakedness relative to the normal distribution. The kurtosis is lower than 3 indicates light tails and flatness [20, 21].

15.5 Result and Discussion

There are approximately thirty-first heart sounds (S1) and 30 s heart sounds (S2) component in 30 s recording of heart sound of each patient. The skewness and kurtosis of this entire S1 and S2 component were calculated for all recorded heart sounds. Table 15.2 shows mean and standard deviation (std.) of the skewness and kurtosis of the heart sounds of one patient with mechanical heart valve thrombosis (Thr), the heart sounds of the same patients after thrombolytic treatment (AThr) and the heart sounds of five patients with normally functioning mechanical heart valve (N). Figure 15.2 illustrates the summary statistics for the skewness and kurtosis of S1 and S2 of Thr, AThr and N.

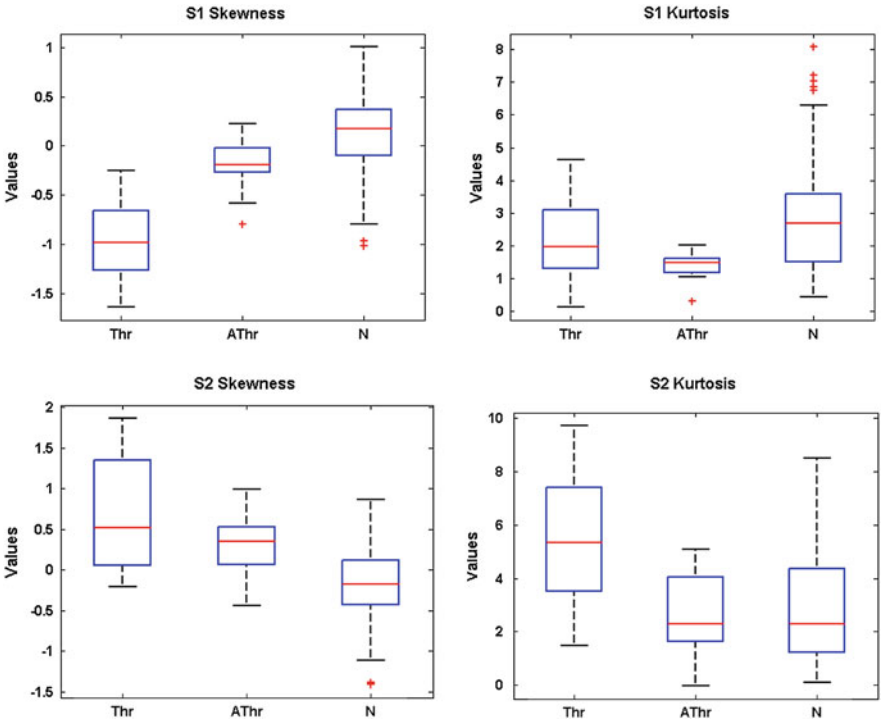


Fig. 15.2 Box plot for the skewness of S1, the skewness of S2, the kurtosis of S2, the kurtosis S1 (Thr: patient with mechanical valve thrombosis, AThr: patient after thrombolytic treatment and N: five patients with normally functioning mechanical valve)

From Table 15.3 and Fig. 15.2, it can be said that there are a meaningful differences between means of the skewness of S1 and S2 and S2 of the kurtosis of heart sounds of patients with normally and malfunctioning mechanical heart valves. The kurtosis of S1 has the similar mean for these heart sounds. However, it is clearly seen that the skewness of S1 is the best feature to show difference between the normally and malfunctioning mechanical heart valve.

Paired *t*-test with 99% confidence level was used for comparison means of the skewness and kurtosis between heart sounds of a patient before and after thrombolytic treatment was administered. Unpaired *t*-test with 99% confidence was used for comparison means of the skewness and kurtosis between patients with normally functioning mechanical heart valve and patient with mechanical heart valve thrombosis (before treatment). These tests were applied to two features, the skewness and kurtosis, obtained from two sound components S1 and S2. p value obtained from above tests is shown on Table 15.3.

The skewness of S2 only between Thr and N, the kurtosis of S1 only between Thr and AThr, the kurtosis of S2 only between Thr and N shows statistically significance differences ($p < 0.01$). But the skewness of S1 shows statistically

Table 15.3 *p* value obtained from paired and unpaired *t*-test

	Between heart sounds of Thr and AThr (paired *t*-test)	Between heart sounds of Thr and N (unpaired *t*-test)
Skewness of S1	0.0000018	0.000006
Skewness of S2	0.0199	0.0011
Kurtosis of S1	0.0046	0.0712
Kurtosis of S2	0.0016	0.0924

Thr Patient with mechanical heart valve thrombosis
AThr Patient with mechanical heart valve thrombosis after thrombolytic treatment
N Five patients with normally functioning mechanical heart valve

significance differences both between Thr and AThr and between Thr and N ($p < 0.01$). Because of this, the skewness of S1 is the best feature to distinguish heart sounds of a patient with mechanical valve thrombosis and normally functioning mechanical heart valve.

15.6 Conclusion and Future Work

In this chapter, the skewness and kurtosis of heart sounds of patients with mechanical heart valve thrombosis and normally functioning mechanical heart valve were compared statistically. As a result, the skewness of S1 of mechanical heart valve should perform fairly well in differentiating normally functioning and malfunctioning mechanical heart valve. However, effectiveness of the skewness of S1 to detect malfunctioning mechanical heart valve should be proven with a large patient population. After that, the skewness of S1 of mechanical heart sound signals may be used for analysis of mechanical heart valve sounds with a view to detecting thrombosis formation on mechanical heart valve.

Acknowledgments This work was supported by scientific research projects (BAP) coordinating office of Selçuk University.

References

1. Edmunds LH, Clark RE, Cohn LH, Grunkemeier GL, Miller DC, Weisel RD (1996) Guidelines for reporting morbidity and mortality after cardiac valvular operations. J Thorac Cardiovasc Surg 112:708–711
2. Roudaut R, Lafitte S, Roudaut MF, Courtault C, Perron JM, Jaı C et al (2003) Fibrinolysis of mechanical prosthetic valve thrombosis. J Am Coll Cardiol 41(4):653–658
3. Caceres-Loriga FM, Perez-Lopez H, Santos-Gracia J, Morlans-Hernandez K (2006) Prosthetic heart valve thrombosis: pathogenesis, diagnosis and management. Int J Cardiol 110:1–6
4. Roscıtano A, Capuano F, Tonellı E, Sınatra R (2005) Acute dysfunction from thrombosis of mechanical mitral valve prosthesis. Braz J Cardiovasc Surg 20(1):88–90

5. Schlitt A, Hauroeder B, Buerke M, Peetz D, Victor A, Hundt F, Bickel C et al (2002) Effects of combined therapy of clopidogrel and aspirin in preventing thrombosis formation on mechanical heart valves in an ex vivo rabbit model. Thromb Res 107:39–43
6. Koller PT, Arom KV (1995) Thrombolytic therapy of left-sided prosthetic valve thrombosis. Chest 108:1683–1689
7. Kaymaz C, Özdemir N, Çevik C, Izgi C, Özveren O, Kaynak E et al (2003) Effect of paravalvular mitral regurgitation on left atrial thrombosis formation in patients with mechanical mitral valves. Am J Cardiol 92:102–105
8. Kim SH, Lee HJ, Huh JM, Chang BC (1998) Spectral analysis of heart valve sound for detection of prosthetic heart valve diseases. Yonsei Med J 39(4):302–308
9. Kim SH, Chang BC, Tack G, Huh JM, Kang MS, Cho BK, Park YH (1994) In vitro sound spectral analysis of prosthetic heart valves by mock circulatory system. Yonsei Med J 35(3):271–278
10. Candy JV, ve Meyer AW (2001) Processing of prosthetic heart valve sounds from anechois tank measurements. 8. International Congress on Sound and Vibration. China
11. Grigioni M, Daniele C, Gaudio CD, Morbiducci U, D'avenio G, Meo DD, Barbaro V (2007) Beat to beat analysis of mechanical heart valves by means of return map. J Med Eng Technol 31(2):94–100
12. Sava HP, Bedi R, McDonnell TE (1995) Spectral analysis of carpentier-edwards prosthetic heart valve sounds in the aortic position using svd-based methods. Signal Process Cardiogr IEE Colloq 6:1–4
13. Sava HP, McDonnell JTE (1996) Spectral composition of heart sounds before and after mechanical heart valve imdantation using a modified forward-backwar d Prony's method. IEEE Trans Biomed Eng 43(7):734–742
14. Altunkaya S, Kara S, Görmüş N, Herdem S (2010) Statistically evaluation of mechanical heart valve thrombosis using heart sounds. Lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering 2010, WCE 2010, 30 June–2 July, 2010, London, UK, 704–708
15. Pan J, Tompkins WJ (1985) A real-time QRS detection algorithm. IEEE Trans Biomed Eng 32(3):230–236
16. Köhler BU, Hennig C, Orglmeister R (2002) The principles of software QRS detection. IEEE Eng Med Biol Mag 21(2):42–57
17. Syed Z, Leeds D, Curtis D, Nesta F, Levine RA, Guttag J (2007) A framework for the analysis of acoustical cardiac signals. IEEE Trans Biomed Eng 54(4):651–662
18. Choi S, Jiang Z (2008) Comparison of envelope extraction algorithms for cardiac sound signal segmentation. Expert Syst Appl 34(2):1056–1069
19. El-Segaier M, Lilja O, Lukkarinen S, Ornmo LS, Sepponen R, Pesonen E (2005) Computer-based detection and analysis of heart sound and murmur. Ann Biomed Eng 33(7):937–942
20. Sanei S, Chambers JA (2007) EEG signal processing. Wiley, Chichester, pp 50–52
21. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1993) Numerical recipes in C: the art of scientific computing. Cambridge University Press, Cambridge, pp 610–615

Chapter 16
On the Implementation of Dependable Real-Time Systems with Non-Preemptive EDF

Michael Short

Abstract Non-preemptive schedulers remain a very popular choice for practitioners of resource constrained real-time embedded systems. This chapter is concerned with the non-preemptive version of the Earliest Deadline First algorithm (npEDF). Although several key results indicate that npEDF should be considered a viable choice for use in resource-constrained real-time systems, these systems have traditionally been implemented using static, table-driven approaches such as the 'cyclic executive'. This is perhaps due to several popular misconceptions regarding the basic operation, optimality and robustness of this algorithm. This chapter will attempt to redress this balance by showing that many of the supposed 'problems' attributed to npEDF can be easily overcome by adopting appropriate implementation and analysis techniques. Examples are given to highlight the fact that npEDF generally outperforms other non-preemptive software architectures when scheduling periodic and sporadic tasks. The chapter concludes with the observation that npEDF should in fact be considered as the algorithm of choice for such systems.

16.1 Introduction

This chapter is concerned with the non-preemptive scheduling of recurring (periodic/sporadic) task models, with applications to resource-constrained, single-processor real-time and embedded systems. In particular, it is concerned with scheduler architectures, consisting of a small amount of hardware (typically a timer/interrupt controller) and associated software. In this context, there are two

M. Short (✉)
Electronics & Control Group, Teesside University, Middlesbrough, TS1 3BA, UK
e-mail: m.short@tees.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_16,
© Springer Science+Business Media B.V. 2011

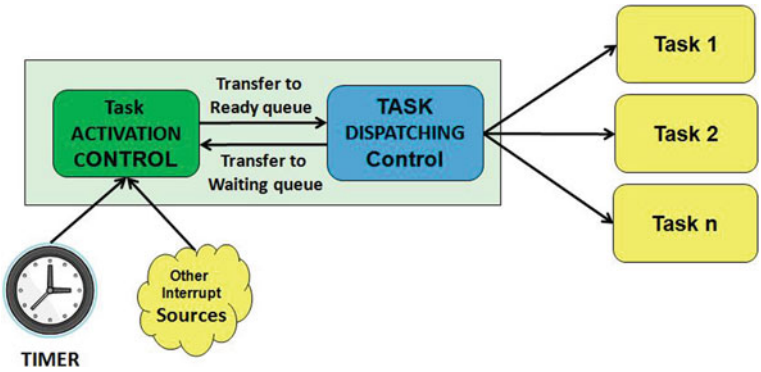


Fig. 16.1 Aspects of real-time embedded scheduling

main requirements of a scheduler. The first is *task activation*, which is the process of deciding at which points in time tasks become ready for execution. Periodic tasks are normally activated via a timer; event driven (sporadic) tasks can be either directly activated by interrupts or by polling an interrupt status flag. The second is *task dispatching*, which is the process of deciding which of the active tasks is best to execute, and some form of scheduling algorithm is normally required to achieve this. These two main aspects of scheduling are illustrated in Fig. 16.1.

The performance of scheduling algorithms and techniques is an area worthy of study; the seminal paper of Liu and Layland [1], published in 1973, spawned a multitude of research and a significant body of results can now be found in the literature. Liu and Layland were the first to discuss deadline-driven scheduling techniques in the context of real-time and embedded computing. It is known that when task preemption is allowed, this technique—also known as Earliest Deadline First (EDF)—allows the full utilization of the CPU, and is optimal on a single processor under a wide variety of different operating constraints [1–3]. However, for developers of systems with severe resource constraints, preemptive scheduling techniques may not be viable; the study of non-preemptive alternatives is justified for the following (non-exhaustive) list of reasons [4–7]:

- Non-preemptive scheduling algorithms are easier to implement than their preemptive counterparts, and can exhibit dramatically lower runtime overheads;
- Non-preemptive scheduling naturally guarantees exclusive access to resources, eliminating the need for complex resource access protocols;
- Task sets under non-preemptive scheduling can share a common stack and processor context, leading to vastly reduced memory requirements;
- Cache and pipeline-related flushing following task preemptions does not occur in non-preemptive systems;
- Implementation of overload detection and recovery methods can be easier to implement;
- Initial studies seem to indicate that non-preemptive systems are less susceptible to transient errors than their preemptive counterparts.

Along with these advantages, non-preemptive scheduling is also known to have several associated disadvantages; task response times are generally longer, event-driven (sporadic) task executions are not as well supported, and when preemption is not allowed, many scheduling problems become either NP-Complete or NP-Hard [8]. This work is concerned with systems implementing the non-preemptive version of EDF (npEDF). The main motivating factors for the work are as follows. Although the treatment of npEDF has been (comparatively) small in the literature, several key results exist that indicate npEDF can overcome most (perhaps not all) of the problems associated with non-preemption; as such it should be considered as a viable choice for use in resource-constrained real-time and embedded systems. However, such systems have traditionally been implemented using static, table-driven approaches such as the 'cyclic executive' and its variants (see, for example, [4, 9–11]). This is perhaps due to several popular misconceptions[1] with respect to the basic operation, implementation complexity, optimality and robustness of the npEDF algorithm, leading to a general lack of coverage in the wider academic community.

This chapter will attempt to redress this balance by arguing the case for npEDF, and showing that the supposed 'problems' commonly attributed to it either simply do not hold, or can easily be overcome by adopting an appropriate implementation and by applying simple off-line analysis techniques. The chapter is organized as follows. Section 16.2 considers why npEDF seems to be 'missing' from most major texts on real-time systems. Section 16.3 presents the assumed task model, and gives a basic description of npEDF. This section also identifies and expands a list of its common criticisms. Section 16.4 subsequently addresses each of these criticisms in turn, to establish their validity (or otherwise). Section 16.5 concludes the chapter.

16.2 npEDF: A Missing Algorithm?

In most of the major texts in the field of real-time systems, npEDF does not get more than a passing mention. For example, analysis of non-preemptive scheduling is typically restricted to the use of 'cyclic executives' or 'timeline schedulers'. In almost all cases, after problems have been identified with such scheduling models, attention is then focused directly on Priority-Driven Preemptive (PDP) approaches as a 'cure for all ills'. For example, Buttazzo [5] discusses timeline scheduling in C4 of his (generally) well-respected book on hard real-time computing systems, concluding with a list of problems associated with this type of scheduling. On p78—immediately before moving onto descriptions of PDP algorithms—it is stated that "The problems outlined above of timeline scheduling can be solved by using priority-based [preemptive] algorithms."

[1] The key results for npEDF—and their implications—are comparatively more difficult to interpret that for other types of scheduling; for example, many previous works assume the reader possesses an in-depth understanding of formal topics in computer science, such as computational complexity.

Liu takes a similar approach in what is perhaps the most widely-acclaimed book in this area (Real-Time Systems) [12]. Cyclic scheduling is discussed in C5 of her book, ending with a list of associated problems on p122. In each case, it is stated that a PDP system can overcome the problem. This type of argument is by no means limited to reference texts. Burns et al. [9] describe (in-depth) some techniques that can be used for generating feasible cyclic or timeline schedules, followed by a discussion of the problems associated with this type of scheduling, directly followed by a final section (p160) discussing "Priority [-based preemptive] scheduling as an alternative to cyclic scheduling". Whilst it is clearly untrue to say these statements are false, as stated above PDP scheduling is not without its own problems; the next section will examine the basics of npEDF, and examine why it seems to have been overlooked.

16.3 Task Model and Preliminaries

16.3.1 Recurring Computational Tasks

This work is concerned with the implementation of recurring/repeated computations on a single processor, such as those that may be required in signal processing and control applications. Such a system may be represented by a set τ of n tasks, where each task $t_i \in \tau$ is represented by the tuple:

$$t_i = (p_i, c_i, d_i) \tag{16.1}$$

In which p_i is the task period (minimum inter-arrival time), c_i is the (worst-case) computation requirement of the task and d_i is the task (relative) deadline. This model was introduced by Liu and Layland [1] and has since been widely adopted—see, for example, [2–7]. Note that it can be assumed w.l.o.g. that time is discrete, and all task parameters can be assumed to be integer [13]. Although implicit deadline tasks (i.e. those in which $d_i = p_i$) are most commonly discussed in the literature (and employed in practice), no specific relationships between periods and deadlines are assumed to fully generalize the work. Note that periodic tasks may additionally be described by an addition parameter, an initial release time (or offset phasing) r_i. Finally, the system utilization U represents the fraction of time the processor will be occupied processing the jobs in the task set over its lifetime, and is defined as $U = \sum c_i/p_i$.

16.3.2 npEDF Algorithm Operation

The npEDF algorithm may be described, in simple terms, as follows:

1. When selecting a task for execution, the task with the earliest deadline is selected first (and then run to completion).

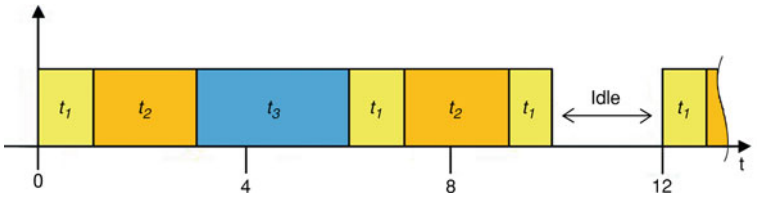


Fig. 16.2 Example schedule generated by npEDF

2. Ties between tasks with identical deadlines are broken by selecting the task with the lowest index.
3. Unless the processor is idle, scheduling decisions are only made at task boundaries.
4. When the scheduler is idle, the first task to be invoked is immediately executed (if multiple tasks are simultaneously invoked, the task with the earliest deadline is selected).

This simple (but deceptively effective) algorithm may be implemented using only a single hardware timer for periodic tasks. The algorithm clearly differs from the static table-driven approaches, in that the schedule is built on-line, and there is therefore no concept of a fixed time 'frame' or 'tick'. An example schedule which is built by npEDF for the set of synchronous tasks τ = [(4, 1, 4), (6, 2, 6), (12, 3, 12)] is shown in Fig. 16.2.

16.3.2.1 npEDF: Common Criticisms

As mentioned in the introduction, generally due to misconceptions (or misinterpretations) of its basic operation and use, npEDF is generally seen to be too problematic for use in real systems. The main criticisms that can be found in the literature are listed below:

1. npEDF is not an optimal non-preemptive scheduling algorithm;
2. npEDF is difficult to analyze, and no efficient schedulability tests exists;
3. npEDF is not 'robust' to changes in the task set parameters;
4. Timer rollover can lead to anomalies and deadline misses in an otherwise schedulable task set;
5. The use of npEDF leads to increased overheads (and power consumption) compared to other non-preemptive scheduling techniques.

Note that optimal in this sense refers to the ability of npEDF to build a valid schedule, if such a schedule exists. Additionally, robustness refers to the ability of a scheduling algorithm to tolerate run-time reductions in the execution requirement of one (or more) tasks (or, equivalently, increases in period) without deadline misses occurring in an otherwise schedulable task set. Please also note that apart from point 3, this list of criticisms is specific to npEDF, and therefore does not

include the so-called 'long-task' problem which is endemic to all non-preemptive schedulers. This specific problem arises when one or more tasks have a deadline that is less than the execution time of another task. In this situation, effective solutions are known to include code-refactoring at the task level, employing state-machines, or alternately adopting the use of hybrid designs [4, 8, 14]. Such solution techniques easily generalize to npEDF, and are not discussed in any further depth here.

16.4 The Case for npEDF

If all of the criticisms given in the previous section are based in fact, then npEDF does not seem a wise choice for system implementation; in fact the contrary would be true. This section will examine each point in greater detail, to investigate if, in fact, each specific claim actually holds.

16.4.1 npEDF is not Optimal

As mentioned previously, optimal in this sense refers to the ability of a scheduling algorithm to build a valid schedule for an arbitrary set of feasible tasks, if such a schedule exists. Each (and every) proof that npEDF is sub-optimal relies on a counter-example of the form shown in Fig. 16.3 (taken from Liu [12]—a similar example appears in Buttazzo [5]). It can be seen that despite the existence of a feasible schedule, obtained via the use of a scheduler which inserts idle-time between $t = 3$ and $t = 4$ (indicated by the question marks in the figure), the schedule produced by npEDF misses a deadline at $t = 12$.

Since the use of inserted idle-time can clearly have a beneficial effect with respect to meeting deadlines, a related question immediately arises: how complex is a scheduler that uses inserted idle time, and will such a scheduler be of practical use for a real system? The answer, unfortunately, is a resounding no. Two important results were formally shown by Howell and Venkatro [15]. The first is that there cannot be an optimal on-line algorithm using inserted idle-time for the non-preemptive scheduling of sporadic tasks; only non-idling scheduling strategies can be optimal. The second is that an on-line scheduling strategy that makes use of inserted idle-time to schedule non-preemptive periodic tasks cannot be efficiently implemented unless $P = NP$. It can thus be seen that inserted idle-time is not beneficial when scheduling sporadic tasks, and if efficiency is taken into account, then attention must be restricted to non-idling strategies when scheduling periodic tasks. Efficiency in this sense refers to the amount of time taken by the scheduler to make scheduling decisions; only schedulers that take time proportional to some polynomial in the task set parameters can be considered efficient (a scheduler which takes 50 years to decide the optimal strategy for the next 10 ms is not much

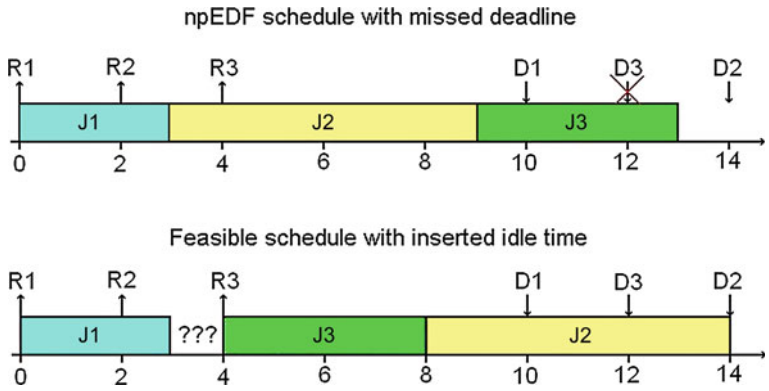


Fig. 16.3 npEDF misses a deadline, yet a feasible schedule exists

practical use). What is known about the non-idling scheduling strategies? These include, for example, npEDF, TTC scheduling [4, 14] and non-preemptive Rate Monotonic (npRM) scheduling [16]. npEDF is known to be optimal among this class of algorithms for scheduling recurring tasks; results in this area were known as early as 1955 [17]. The proof was demonstrated in the real-time context by Jeffay et al. [6] for the implicit deadline case, and extended by George et al. [18] to the constrained deadline case. Thus, the overall claim status: npEDF is sub-optimal for periodic tasks if and only if $P = NP$, and is optimal for sporadic tasks regardless of the equivalence (or otherwise) of these complexity classes.

16.4.2 No Efficient Schedulability Test Exist for npEDF

Consider again the example shown in Fig. 16.3, in which the npEDF algorithm misses a deadline. Why is the deadline missed? At $t = 3$, only J2 is active and, since the scheduler is non-idling, it immediately begins execution of this task. Subsequently at $t = 4$, J3 is released and has an earlier deadline—but it is blocked (due to non-preemption) until J2 has run to completion at $t = 9$. This is known as a 'priority inversion' as the scheduler cannot change its mind, once committed. This is highlighted further in Fig. 16.4.

Worst-case priority inversions under npEDF scheduling have been investigated in some detail. A relatively simple set of conditions for implicit deadline tasks was derived by Jeffay et al. [6], and were subsequently generalized by George et al. to the arbitrary deadline case [18]. They showed that a set of arbitrary-deadline periodic/sporadic tasks is schedulable under npEDF if and only if all deadlines are met over a specific analysis interval (of length L) following a synchronous arrival sequence of the tasks at $t = 0$, with the occurrence of worst-case blocking due to non-preemption simultaneously occurring. This situation is depicted in Fig. 16.5,

Fig. 16.4 Priority inversion due to non-preemption

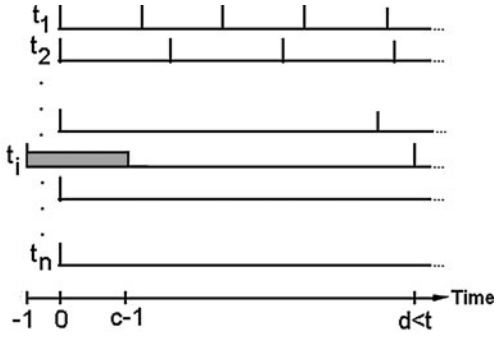


Fig. 16.5 Worst-case priority inversion induced by task i arriving at $t = -1$

showing the task with the largest execution time beginning execution one time unit prior to the simultaneous arrival of the other tasks.

These conditions can be formalized to obtain a schedulability test, which is captured by the following conditions:

$$\begin{aligned} &U \leq 1.0; \\ &\forall t,\ 0 < t < L;\ hb(t) \leq t; \end{aligned} \tag{16.2}$$

Where:

$$hb(t) = \sum_{i=1}^{i=n} \max\left\{0, \left\lfloor \frac{t + p_i - d_i}{p_i} \right\rfloor \cdot c_i\right\} + \max_{d_i > t}\{c_i - 1\} \tag{16.3}$$

And:

$$L = \max\left\{d_1, \ldots, d_n, \frac{\sum_{i=1}^{i=n} (p_i - d_i) \cdot U_i}{1 - U}\right\} \tag{16.4}$$

It should be noted that the time complexity of an algorithm to decide Eqs. 16.2–16.4 is pseudo-polynomial (and hence highly efficient) whenever $U < 1.0$. Other upper bounds on the length on L are derived in [18]. The non-preemptive scheduling problem, in this formulation, turns out to be only weakly coNP-Complete. When compared to feasibility tests for other non-preemptive scheduling disciplines, this is significantly better. For example, it is known that deciding if a

set of periodic process can be scheduled by a cyclic executive or timeline scheduler is strongly NP-Hard [8, 9]; it is also known that deciding if a set of periodic process can be scheduled by a TTC scheduler is strongly coNP-Hard[2] [19]. Note that strong and weak complexity results have a precise technical meaning; specifically, amongst other things the former rules out the prospect of a pseudo-polynomial time algorithm unless $P = NP$, whereas the latter does not. Thus, although a very efficient algorithm may be formulated to exactly test for Eqs. 16.2–16.4, it is thought that no exact algorithm can ever be designed to efficiently test schedulability for these alternate scheduling policies. Overall claim status: *npEDF admits an efficient feasibility test for periodic* (*sporadic*) *tasks that ensures even worst-case priority inversions do not lead to deadline misses.*

16.4.3 npEDF is not Robust to Reductions in System Load

With respect to this complaint, Jane Liu presents some convincing evidence on p. 73 of her book Real-Time Systems [12], and cites the seminal paper by Graham [20] investigating timing anomalies. There are two principal problems here. The paper by Graham deals only with the multiprocessor case; specifically, it investigates the effects of reduced (aperiodic) task execution times on the makespan produced by the LPT heuristic scheduling technique. As do the examples on p. 73 of Liu's book, although it is not made explicitly clear. With respect to single-processor scheduling, these examples simply do not apply; the only single processor timing anomaly referred to in the Liu text is reproduced in Fig. 16.6; at first glance, it seems that a run-time reduction in the execution requirement of job C1 does, indeed, lead to a deadline miss of J3.

However upon closer inspection, this example can be seen to be almost identical to the example given in Fig. 16.3, with the execution of J1 between $t = 3$ and $t = 4$ effectively serving the same purpose as the inserted idle-time in Fig. 16.3. In order for this example to hold up, it must logically follow that the schedule must be provably schedulable when the tasks have nominal parameters given by A); applying Eqs. 16.2–16.4 to these tasks, it can be quickly determined that the task set is not deemed to be schedulable, since the formulation of Jeffay's feasibility test takes worst-case priority inversion into account.

This example is misleading w.r.t. npEDF—since the task set simply fails the basic feasibility test, Liu's argument of 'an otherwise schedulable task set' becomes a non-starter. This again highlights the fact that misconceptions regarding robustness and priority inversions have principally arisen from one simple fact; as shown in the previous section, the worst case behavior of a task set—its critical

[2] In fact, this situation is known to considerably worse than this. The problem is actually known to be NP^{NP}-Complete [19]. Under the assumption that $P \neq NP$, this means that the feasibility test requires an exponential number of calls to a decision procedure which is itself strongly coNP-Complete.

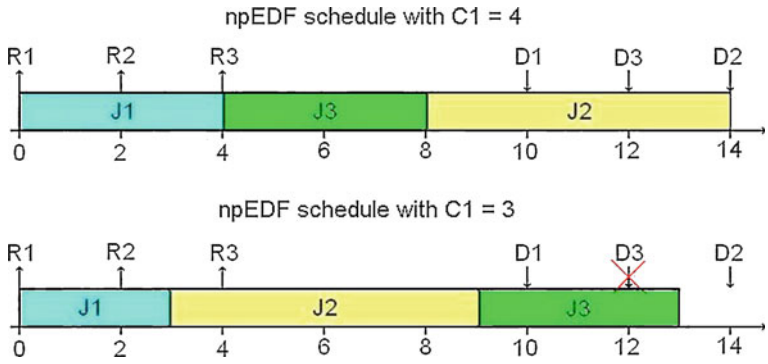


Fig. 16.6 Evidence for a lack of npEDF robustness?

Fig. 16.7 Converting from an absolute (*left*) to a modular (*right*) representation of time

```
// Timestamp variables
uint32_t x, y;

// Timestamp comparison
if(x < y)
  {
  // Conditional Code
  }
```

```
// Timestamp variables
uint32_t x, y;

// Timestamp comparison
if(signed(x - y)<0)
  {
  // Conditional code
  }
```

instants—under non-preemptive scheduling is not the same as under preemptive scheduling. Overall claim status: If appropriate (off-line) analysis is performed to confirm the schedulability of a task set, this task set will remain schedulable under npEDF even under conditions of reduced system load.

16.4.4 Timer Rollover Can Lead to Deadline Misses

With respect to this complaint, this can in fact be shown to hold, but is easily solved. The assumption that time is represented as integer—and in embedded systems, normally with a fixed number of bits (e.g. 16)—eventually leads to timer rollover problems; deadlines will naturally 'wrap around' due to the modular representation of time. Since the normal laws of arithmetic no longer hold, it cannot be guaranteed that $d_i \bmod(2^b) < d_j \bmod(2^b)$ when $d_i < d_j$ and time is represented by b-bit unsigned integers. There are several efficient techniques that may be used to overcome this problem, perhaps the most efficient is as follows. Assuming that the inequality $p_m < 2^b/2$ holds over a given task set, i.e. the maximum period is less than half the linear life time of the underlying timer, then the rollover problem may be efficiently overcome by using Carlini and Buttazzo's Implicit Circular Timer Overflow Handler (ICTOH) algorithm [21]. The algorithm has a very simple code implementation, and is show as C code in Fig. 16.7.

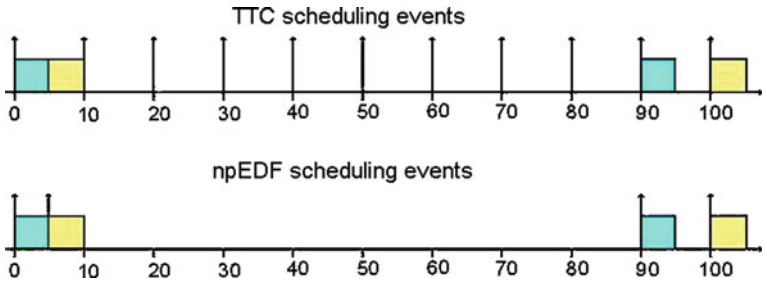


Fig. 16.8 Density of scheduling events for both TTC and npEDF scheduling

The algorithm's operation exploits the fact that the modular distance between any two events (e.g. deadlines or activation times) x and y, encoded by b-bit unsigned integers, may be determined by performing a subtraction modulo 2^b between x and y, with the result interpreted as a signed integer. Overall claim status: rollover is easily handled by employing algorithms such as ICTOH.

16.4.5 npEDF Use Leads to Large Scheduling Overheads

In order to shed more light on this issue, let us consider the required number of 'scheduling events' over the hyperperiod (major cycle) of a given periodic task set, and also the complexity—the required CPU iterations, as a function of the task parameters—of each such event. Specifically, let us consider these scheduling events as required for task sets under both npEDF and TTC scheduling. TTC scheduling is considered as the baseline case in this respect, as it has previously been argued that a TTC scheduler provides a software architecture with minimal overheads and resource requirements [4, 7, 14]. With npEDF, one scheduling event is required for each and every task execution, and the scheduler enters idle mode when all pending tasks are executed. It can be woken by an interrupt set to match the earliest time at which a new task will be invoked. The TTC algorithm is designed to perform a scheduling event at regular intervals, in response to periodic timer interrupts; the period of these interrupts is normally set to be the greatest common divisor of the task periods [4, 14]. Let the major cycle h of a set of synchronous tasks be given by $h = lcm(p_1, p_2, \ldots, p_n)$. The number of scheduling events occurring in h for both the TTC scheduler—SE_{TTC}—and the npEDF scheduler—SE_{EDF}—are given by:

$$SE_{TTC} = \frac{lcm(p_1, p_2, \ldots, p_n)}{gcd(p_1, p_2, \ldots, p_n)};\ SE_{EDF} = \sum_{i \in \tau} \frac{lcm(p_1, p_2, \ldots, p_n)}{p_i}; \qquad (16.5)$$

Clearly $SE_{EDF} \leq SE_{TTC}$ in almost all cases, and an example to highlight this is shown for the task set $\tau = [(90, 5), (100, 5)]$ in Fig. 16.8, where scheduling events

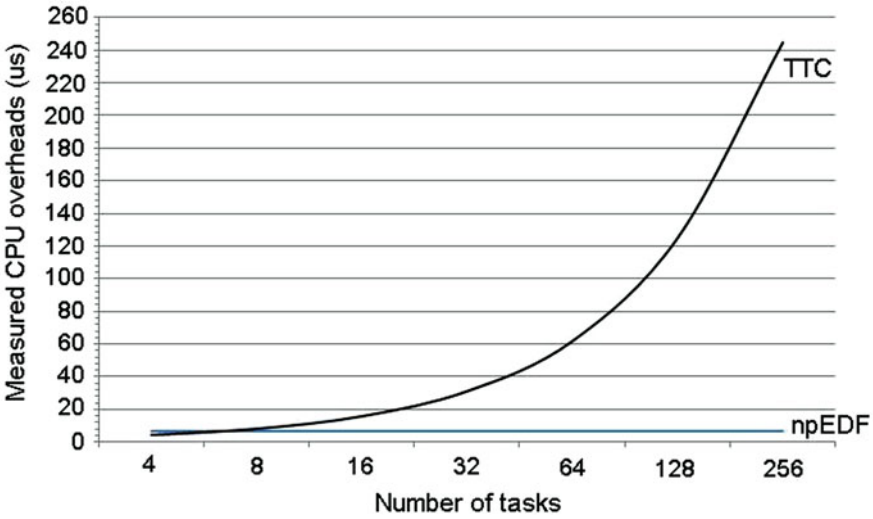


Fig. 16.9 CPU overheads vs. number of tasks

are indicated by the presence of up-arrows on the timeline. Also of interest are the time complexities of each scheduling event. Given the design of the TTC scheduler, it is clear from its implementation (see, for example, [4, 14]) that its complexity is $O(n)$. Task management in the npEDF scheduler significantly improves upon this situation; it is known that the algorithm can be implemented with complexity $O(log\ n)$ or better, in some cases $O(1)$ [22]. To further illustrate this final point, Fig. 16.9 shows a comparison of the overheads incurred per scheduling event as the number of tasks was increased on a 72-Mhz ARM7-TDMI microcontroller. Overhead execution times were extracted using the technique described in [22]. This graph clearly shows the advantages of the npEDF technique, and for $n > 8$ the overheads become increasingly smaller. Overall claim status: *With an appropriate implementation, the density of npEDF scheduling events is significantly better than competing methods; the CPU overheads incurred at each such event are also significantly lower.*

16.5 Conclusions

This chapter has considered the non-preemptive version of the Earliest Deadline First algorithm, and has investigated the supposed problems that have been attributed to this form of scheduling technique. Examples and analysis have been given to show that these problems are either baseless or trivially solved, and in most cases npEDF outperforms many other non-preemptive software architectures. As such, it the conclusion of the current chapter that npEDF should be considered

as one of the primary algorithms for implementing resource-constrained real-time and embedded systems. A preliminary version of the work described in this chapter was presented at the World Congress on Engineering, July 2010 [23].

References

1. Liu J, Layland J (1973) Scheduling algorithms for multiprogramming in a hard real-time environment. J ACM 20(1):46–61
2. Coffman E Jr (1976) Introduction to deterministic scheduling theory, in computer and job-shop scheduling theory. Wiley, New York
3. Dertouzos ML (1974) Control robotics: the procedural control of physical processes. Inf Process 74
4. Pont M (2001) Patterns for time-triggered embedded systems. ACM Press/Addison-Wesley Education, New York
5. Buttazzo GC (2005) Hard real-time computing systems: predictable scheduling algorithms and applications. Spinger, New York
6. Jeffay K, Stanat D, Martel C (1991) On non-preemptive scheduling of periodic and sporadic tasks. In: Proceedings of the IEEE Real-Time Systems Symposium
7. Short M, Pont M, Fang J (2008) Exploring the impact of pre-emption on dependability in time-triggered embedded systems: a pilot study. In: Proceedings of the 20th Euromicro Conference on Real-Time Systems (ECRTS 2008), Prague, Czech Republic, pp 83–91
8. Garey MR, Johnson DS (1979) Computers and intractability: a guide to the theory of NP-completeness. W.H. Freeman & Co Ltd, New York
9. Burns A, Hayes N, Richardson M (1994) Generating feasible cyclic schedules. Control Eng Pract 3(2):151–162
10. Baker TP, Shaw A (1989) The cyclic executive model and Ada. Real-Time Syst 1(1):7–25
11. Locke CD (1992) Software architecture for hard real-time applications, cyclic executives vs. fixed priority executives. Real-Time Syst 4(1):37–52
12. Liu JWS (2000) Real-time systems. Prentice-Hall, New Jersey
13. Baruah S, Rosier L, Howell R (1991) Algorithms and complexity concerning the preemptive scheduling of periodic, real-time tasks on one processor. Real-Time Syst 2(4):301–324
14. Gendy AK, Pont MJ (2008) Automatically configuring time-triggered schedulers for use with resource-constrained, single-processor embedded systems. IEEE Trans Ind Inform 4(1):37–45
15. Howell R, Venkatro M (1995) On non-preemptive scheduling of recurring tasks using inserted idle times. Inf Comput 117:50–62
16. Park M (2007) Non-preemptive fixed priority scheduling of hard real-time periodic tasks. Lect Notes Comput Sci 4990:881–888
17. Jackson JR (1955) Scheduling a production line to minimize maximum tardiness. Research report 43, Management Science Research Project, University of California, Los Angeles, USA
18. George L, Rivierre N, Supri M (1996) Preemptive and non-preemptive real-time uni-processor scheduling. Research report RR-2966, INRIA, Le Chesnay Cedex, France
19. Short M (2009) Some complexity results concerning the non-preemptive 'thrift' cyclic scheduler. In: Proceedings of the 6th International Conference on Informatics in Control, Robotics and Automation (ICINCO 2009), Milan, Italy, July 2009, pp 347–350
20. Graham RL (1969) Bounds on multiprocessing timing anomalies. SIAM J Appl Math 17:416–429
21. Carlini A, Buttazzo GC (2003) An efficient time representation for real-time embedded systems. In: Proceedings of the ACM Symposium on Applied Computing (SAC 2003), Florida, USA, March 2003, pp 705–712

22. Short M (2010) Improved task management techniques for enforcing EDF scheduling on recurring task sets. In: Proceedings of the 16th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS 2010), Stockholm, Sweden, April 2010, pp 56–65
23. Short M (2010) The case for non-preemptive, deadline-driven scheduling in real-time embedded systems. In: Lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering 2010 (WCE 2010), vol 1. London, UK, 30 June–2 July 2010, pp 399–404

Chapter 17
Towards Linking Islands of Information Within Construction Projects Utilizing RF Technologies

Javad Majrouhi Sardroud and Mukesh Limbachiy

Abstract Modern construction management require real-time and accurate information for sharing among all the parties involved to undertake efficient and effective planning as well as execution of the projects. Research projects conducted during the last decade have concluded that information management is a critical factor in construction project performance and plays an essential role in managing the construction where projects need to be completed within a defined budget and deadline. Recently, wireless sensor technologies have matured and become both technically and economically feasible and viable. This research investigates a framework for integrating the latest innovations in Radio Frequencies (RF) based information management system to automate the task of collecting and sharing of detailed and accurate information in an effective way throughout the actual construction projects. The solution presented here is intended to extend the use of a cost-effective and easy-to-implement system (Radio Frequency Identification (RFID), Global Positioning System (GPS), and Global System for Mobile Communications (GSM)) to facilitate low-cost and network-free solutions for obtaining real-time information and information sharing among the involved participants of the ongoing construction projects such as owner, consultant, and contractor.

J. M. Sardroud (✉)
Faculty of Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran
e-mail: J.Majrouhi@Kingston.ac.uk

M. Limbachiy
School of Civil and Construction Engineering, Kingston University London, Kingston upon Thames, London, KT1 2EE, UK
e-mail: M.Limbachiya@kingston.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_17,
© Springer Science+Business Media B.V. 2011

17.1 Introduction

Construction is identified internationally as one of the information-intensive industry which subject to open environment and survive harsh conditions [1, 2]. Due to the complex, unprepared, and uncontrolled nature of the construction site, not only using of automated advanced tracking and data storage technologies for efficient information management is needed but also construction industry has greatly benefited from technology in rising the speed of information flow, enhancing the efficiency and effectiveness of information communication, and reducing the cost of information transfer [3]. Missing and delayed information access constitutes 50–80% of the problems in construction. One of the major sources of information is the data collected on construction sites. Even though, collection of detailed, accurate and a sufficient volume of information and timely delivery of it is vital for effective construction management, the current situations of on-site information management methods are manually on the human ability using paper and pencil in all parts of the construction phase [4]. Previous observations on construction sites cite that 30–50% of the field supervisory personnel's time is spent on recording and analyzing field data [5] and 2% of the work on construction sites is devoted to manual tracking and recording of progress data [6]. Data collected using manual methods are not reliable or complete due to reluctance of workers to monitor and record the flow of large quantities of elements. Data collected through these methods are usually transferred and stored in paper-based format, which is difficult to search and access, and makes processing data into useful information expensive and unreliable. Thus, some information items end up being unavailable to the parties who need access to them in a timely manner to make effective decisions [7]. Effective and immediate access to information minimizes the time and labour used for retrieving information related to each part of construction and reduces the occurrence of ineffective decisions that are made in the absence of information [8]. The process of capturing quantity of work data at a construction site needed to be improved in terms of accuracy and completeness to eliminate unnecessary communication loops and secondary tasks caused by missing or inaccurate data. These all suggest the need for a fully automatic data collection technology to capture the status information throughout construction and to integrate this data in a database automatically. The emergence of ubiquitous system which is developed in this research has the potential to enlarge the boundary of information systems from the actual work sites to the site offices and ensure real time data flow among all participants of construction projects.

This research investigates the fully automated data collection using integrated applications of Radio Frequency Identification (RFID), Global Positioning System (GPS) and Global System for Mobile Communication (GSM) technologies in the construction industry which focused on the real-time exchange of information between the on and off construction sites. The system addresses a clear path for obtaining real-time information and information sharing among the involved

participants of the construction phase such as owner, consultant, and contractor via the Internet. The solution presented here is intended to extend the use of current technologies RFID, GPS, and GSM to facilitate extremely low-cost and network-free solutions to form the backbone of an information management system for practical communication and control among construction participants. The remainder of this paper first reviews previous research efforts that have been done by others relating to applications of wireless technologies in construction, followed by an assessment of the enabling technologies which are utilized in this research. Then it reveals the architecture of the integrated system for collecting and sharing real time information in all part of construction phases. Finally conclusions are given in the end.

17.2 Prior Research Efforts

Many research projects have focused on the application of wireless technologies in the construction sector [9]. These technologies can be used for tracking, locating, and identifying materials, vehicles and equipment that lead to important changes in managerial tasks in the construction industry [10]. Recent research projects looked on the potential of using wireless technologies in the construction sector to improve the process of capturing data [11–13], some of these are discussed here. Jaselskis et al. have summarized RFID technology and surveyed its potential applications in the construction industry including concrete processing and handling, cost coding of labour and equipment, and material control [14]. Chen et al. conducted an experiment in which Bar-code technique is used to facilitate effective management of construction materials to reduce construction wastes [15]. Jaselskis and El-Misalami implemented RFID to receive and keep tracking of pipe spool in the industrial construction process. Their pilot test demonstrated that RFID could increase operation efficiency by saving time and cost in material receiving and tracking [16]. Oloufa et al. examined the use of differential GPS technology on construction sites to avoiding equipment collisions [17]. Jang and Skibniewski developed an Automated Material Tracking system based on ZigBee localization technology with two different types of query and response pulses [18]. Song et al. developed a system that can identify logistics flow and location of construction materials with better performance by using wireless sensor networks such as ZigBee technologies [19]. Majrouhi Sardroud and Limbachiya investigated the use of RFID technology in construction information delivery and management [4]. In some research efforts, authors have developed RFID based methods to automate the task of tracking and locating of construction materials and components in lay down yards and under shipping portals [20–29] and to improve the efficiency of tracking tools and movement of construction equipment and vehicles on and off construction sites [30–34].

Although, the aforementioned research has proven the value and potential of using wireless technologies, however, the reality is that the use of a cost-effective,

scalable, and easy-to-implement information management system in an effective way at actual construction projects are scarce. This research created a framework for integrating the latest innovations in Automated Data Collection (ADC) technologies such as RFID, GPS, and GSM that address a clear path to automate collecting and sharing of detailed, accurate and a sufficient volume of information throughout the construction phase using minimal or no human efforts.

17.3 Technology Description

Recently, wireless sensor technologies have matured and become both technically and economically feasible and viable to potentially support information delivery and management for construction industry. Advanced tracking and data storage technologies such as RFID, GPS, and GSM provide an automated data collection on construction phases and allow all participants to share data accurately, completely, and almost instantly.

In recent years, RFID attracts attention as an alternative to the bar code and has been successfully applied to the areas of manufacturing, distribution industry, supply chain, agriculture, transportation, and healthcare [35, 36]. RFID is a method of remotely storing and retrieving data by utilizing radio frequency in identifying, tracking, and detecting various objects [37]. An early, if not the first, work exploring RFID is the landmark paper by Harry Stockman, "Communication by Means of Reflected Power" [38]. A RIFD system consists of tags (transponder) with an antenna, a reader (transceiver) with an antenna, and a host terminal. The RFID reader acts as a transmitter/receiver and transmits an electromagnetic field that ''wakes-up'' the tag and provides the power required for the tag to operate [3]. A typical RFID system is shown in Fig. 17.1. An RFID tag is a portable memory device located on a chip that is encapsulated in a protective shell and can be attached to any object which stores dynamic information about the object. Tags consist of a small integrated circuit chip coupled with an antenna to enable them to receive and respond to radio frequency queries from a reader. Tags can be

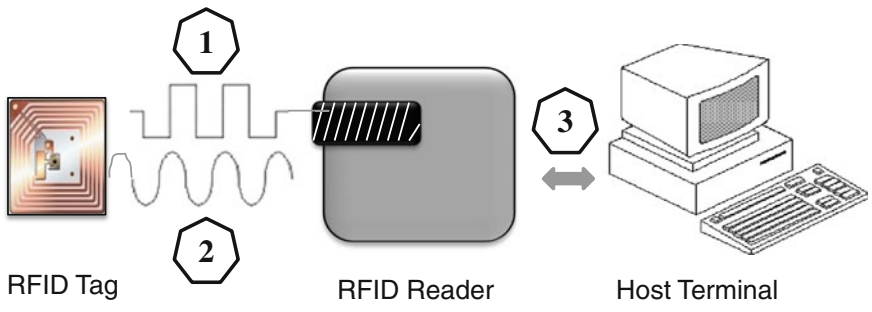


Fig. 17.1 A typical RFID system

categorized as read-only (RO), write once, read many (WORM), and read-write (RW) in which the volume capacity of their built-in memories varies from a few bits to thousands of bits. RFID tags can be classified into active tags (battery powered) and passive tags, which powered solely by the magnetic field emanated from the reader and hence have an unlimited lifetime.

Reading and writing ranges are depend on the operation frequency (low, high, ultra high, and microwave). Low frequency systems generally operate at 124, 125 or 135 kHz. High frequency systems operates at 13.56 MHz and ultra high frequency (UHF) and use a band anywhere from 400 to 960 MHz [39]. Tags operating at ultra high frequency (UHF) typically have longer reading ranges than tags operating at other frequencies. Similarly, active tags have typically longer reading ranges than passive tags. Tags also vary by the amount of information they can hold, life expectancy, recycle ability, attachment method, usability, and cost. Communication distance between RFID tags and readers may decrease significantly due to interferences by steel objects and moisture in the vicinity, which is commonplace in a construction site. Active tags have internal battery source and therefore have shorter lifetime of approximately three to ten years [16]. The reader, combined with an external antenna, reads/writes data from/to a tag via radio frequency and transfers data to a host computer. The reader can be configured either as a handheld or a fixed mount device [40]. The host and software system is an all-encompassing term for the hardware and software component that is separate from the RFID hardware (i.e., reader and tag); the system is composed of the following four main components: Edge interface/system, Middleware, Enterprise back-end interface, and Enterprise back end [14]. RFID tags are more durable and suitable for a construction site environment in comparison with Barcodes which are easily peeled off and may be illegible when they become dirty. RFID tags are not damaged as easily and do not require line-of sight for reading and writing, they can also be read in direct sunlight and survive harsh conditions, reusable, and permit remote [4]. According to the shape of assets, RFID tag can be manufactured all kinds of shapes to adapt all kinds of assets [41].

GPS is a Global Positioning System based on satellite technology. The activities on GPS were initiated by the US Department of Defence (DOD) in the early 1970s under the term Navigation Satellite Timing and Ranging System (NAVSTAR). Glonass, Galileo, and BeiDou are Russian, European Union, and Chinese Global Positioning Systems, respectively [42, 43]. GPS consists of nominally 24 satellites that provide the ranging signals and data messages to the user equipment [44]. To calculate locations, the readings from at least four satellites are necessary, because there are four parameters to calculate: three location variables and the receiver's time [45]. To get metric or sub metric accuracy in positioning data (i.e. longitude, latitude, and altitude), a single GPS receiver is not sufficient; instead a pair of receivers perform measurements with common satellites and operate in a differential mode. DGPS provides sufficient accuracy for most outdoor tracking applications. In DGPS two receivers are used. One receiver measures the coordinates of a stationary point, called the base, whose position is perfectly known in the reference geodetic system used by GPS. The 3-D deviation between the

measured and actual position of the base, which is roughly equal to the measurement error at a second receiver at an unknown point (called ‘‘rover’’), is used to correct the position computed by the latter [46].

GSM is a worldwide standard for cellular communications. The idea of cell-based mobile radio systems appeared at Bell Laboratories in the early 1970s. In 1982 the Conference of European Posts and Telecommunications formed the Groupe Spécial Mobile (GSM) to develop a pan-European mobile cellular radio system (the acronym later became Global System for Mobile communications). One of the current available technologies for mobile data transfer is General Packet Radio Systems (GPRS). GPRS is a packet switched “always on” technology which allows data to be sent and received across a mobile telephone network almost instantly, so immediacy is one of the advantages of GPRS [47].

17.4 Architecture of the Proposed System

The RFID-based ubiquitous system (U-Box) utilized in this research is combination of GPS, RFID and GSM, and as such, takes advantage of the respective strengths of each. The system could be divided into two major parts, mobile system and central station. The Mobile system mainly consists of three types of hardware components; namely, (i) GPS technology; (ii) RFID technology where passive High Frequency (HF) and Ultra High Frequency (UHF) band RFID tags is used for identifying and obtaining the object/user related information by using an RFID reader which is plugged into the mobile system; and (iii) GSM communication technology where the information (ID, specific information and date) retrieved from RFID readers and GPS is transferred to the server using GPRS or SMS. The central station consists of two servers, the application server (portal system) and the database server (project database). In this approach, data collection is done continuously and autonomously, therefore, the RFID as a promising technology is the solution for the information collection problems and the portal with GSM technology is used to solve the information communication problem in the construction industry. A schematic model of U-Box is shown in Fig. 17.2.

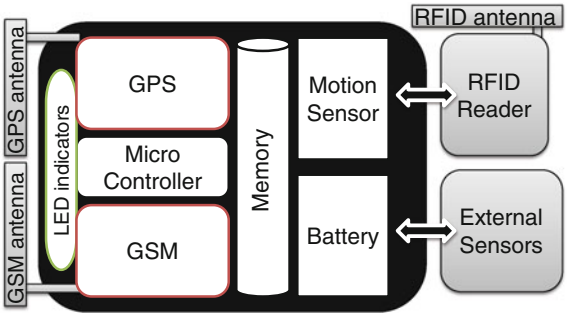


Fig. 17.2 A schematic model of U-Box

As it can be seen in the block diagram, the device has a rechargeable internal battery and a motion sensor. Micro controller checks the source of power. If it's still connected, micro controller sends a command to controller module to recharging the battery. Also, it has its own internal memory to store information (Lat, Long, Data and Time, and etc.), ability to save information when it lost GSM network, and sending saved information immediately after registering in the network. Users can attach some external sensors to U-Box so micro controller will get information of sensors and store them in internal memory then will send to data centre via defined link.

In identification segment, selected RFID technology for any moving probes is active RFID where an independent power supply active RFID tags allow greater communication ranges, higher carrier frequencies, greater data transmission rates, better noise immunity, and larger data storage capacity. In positioning segment the GPS unite will be durable to function in open air conditions. The GPS receiver had a nominal accuracy of 5 m with Wide Area Augmentation System (WAAS). In data transmission segment GPRS and SMS has been selected to support the data transmission between the U-Box and the central office. GPRS connected to the GSM network via SIM card for data transfer enables several new applications which have not previously been available over GSM networks due to the limitations in message length of SMS (160 characters) such as Multimedia Messaging.

In this approach, on site data collection begins with RFID tags that contain unique ID numbers and carries data on its internal memory about the host such as item specific information. It can be placed on any object/user such as materials or workers. During the construction process and at the times of moving any object, the information on the RFID tag is captured and deciphered by the RFID reader which is connected to the mobile system and indeed the micro controller gets information of GPS (which is part of U-Box) and stores the location of the object/user. The ID and location information of the object/user is then sent to a database using GSM technology. In this approach, the tags are used only for identification, and all of the related information is uploaded and stored in one or more databases which will be indexed with the same unique ID of objects. In another mechanism, information can be stored directly on the RFID tags locally and not to store any data in the server. Information update and announcement is synchronously sent via the portal and the system will effectively increase the accuracy and speed of data entry by providing owners, consultants, and contractors with the real time related information of any object/user. The application server defines various applications for collecting, sharing, and managing information. Any moving probes, such as materials handling equipment (top-slewing and bottom-slewing tower cranes, truck-mounted mobile cranes, and crawler), hoists, internal and external delivery vehicle, the gates and some key workers should be equipped with the U-Box. This intelligent system could be programmed to send back information via SMS when RFID reader or user defined sensors which are connected to the system receive new data, for example from uploaded component to the truck or detected data by sensors.

Collected data will be used in the application side by using a web-based portal system for information sharing among all participants. Electronic exchange of

collected information leads to reduction of errors and improved efficiency of the operation processes. The portal system provides an organization with a single, integrated database, both within the organization and among the organizations and their major partners. With the portal system and its coupled tools, managers and workers of each participant can conduct valuable monitor and controlling activities throughout the construction project. For instance, information is transmitted back to the engineering office for analysis and records, enables the generating of reports on productivity where this up-to-date information about construction enables effective management of project.

One of the challenges of designing an effective construction information management system is designing an effective construction tagging system. Each RFID tag is equipped with a unique electronic identity code which is usually the base of reports that contain tracking information for a particular user/object. In choosing the right RFID tag for any application, there are a number of considerations, including: frequency range, memory size, range performance, form factor, environmental conditions, and standards compliance. To minimize the performance reduction of selected technology in contact with metal and concrete, RFID tags need to be encapsulated or insulated. Extremely heavy foliage or underground places like tunnels would cause the signal to fade to an extent when it can no longer be heard by the GPS or GSM antenna. When this happen, the receiver will no longer know its location and the in the case of an intelligent system application, the vehicle is technically lost and central office won't receive information from this system. In this case to locate vehicles inside GPS blind areas, intelligent system will use RFID reader to save tag-IDs in the way through the tunnel—each tag-id shows a unique location—the device will store all information inside internal memory as a current position, and the system will send unsent data to central office when network re-established. In this research, a geo-referenced map of the construction job site should be created once, and then it will be used to identify locations of the objects/users by comparing the coordinates received from the GPS with those in the geo-referenced map.

17.5 Conclusions

Proposed system is an application framework of RFID-based automated data collection technologies which focuses on the real-time collection and exchange of information among the all participants of construction project, construction site and off-site office. This system can provide low-cost, timely, and faster information flow with greater accuracy by using RFID technology, GPS, GSM, and a portal system. In this research data collection is done continuously and autonomously, therefore, the combination of selected Radio Frequencies (RF) based information and communication technologies as a powerful portable data collection tool enables collecting, storing, sharing, and reusing field data accurately, completely, and almost instantly. In this manner up-to-date information

regarding all parts of construction phase is available which permits real-time control enabling corrective actions to be taken. The system enables collected information to be shared among the involved participants of the construction phase via the Internet which leads to important changes in the construction project control and management. The proposed system has numerous advantages. It is automatic, thus reducing the labour costs and eliminating human error associated with data collection during the processes of construction. It can dramatically improve the construction management activities which also lead to keep cost and time under control in the construction phase. The authors believe that, in practice, the approached pervasive system can deliver a complete return on investment within a short period by reducing operational costs and increasing workforce productivity.

References

1. Bowden S, Dorr A, Thorpe T, Anumba C (2006) Mobile ICT support for construction process improvement. Autom Constr 15(5):664–676
2. Behzadan H, Aziz Z, Anumba CJ, Kamat VR (2008) Ubiquitous location tracking for context-specific information delivery on construction sites. Autom Constr 17(6):737–748
3. Wang LC, Lin YC, Lin PH (2007) Dynamic mobile RFID-based supply chain control and management system in construction. Adv Eng Inform 21(4):377–390
4. Majrouhi Sardroud J, Limbachiya MC (2010) Effective information delivery at construction phase with integrated application of RFID, GPS and GSM technology. Lect Notes Eng Comput Sci 2183(1):425–431
5. McCullouch B (1997) Automating field data collection on construction organizations. In: 5th Construction Congress: Managing Engineered Construction in Expanding Global Markets, Minneapolis, USA
6. Cheok GS, Lipman RR, Witzgall C, Bernal J, Stone WC (2000) Non-intrusive scanning technology for construction status determination. Building and Fire Research Laboratory, National Institute of Standards and Technology, NIST Construction Automation Program Report no. 4
7. Ergen E, Akinci B, Sacks R (2003) Formalization and automation of effective tracking and locating of precast components in a storage yard. In: 9th EuropIA International Conference (EIA-9), E-Activities and Intelligent Support in Design and the Built Environment, Istanbul, Turkey
8. Akinci B, Kiziltas S, Ergen E, Karaesmen IZ, Keceli F (2006) Modeling and analyzing the impact of technology on data capture and transfer processes at construction sites: a case study. J Constr Eng Manag 132(11):1148–1157
9. Majrouhi Sardroud J, Limbachiya MC, Saremi AA (2009) An overview of RFID applications in construction industry. In: Third International RFID Conference, 15–16 August, 2009, Tehran, Iran
10. Majrouhi Sardroud J, Limbachiya MC, Saremi AA (2010) Ubiquitous tracking and locating of construction resource using GIS and RFID. In: 6th GIS Conference & Exhibition, (GIS 88), 6 January 2010, Tehran, Iran
11. Pradhan A, Ergen E, Akinci B (2009) Technological assessment of radio frequency identification technology for indoor localization. J Comput Civ Eng 23(4):230–238
12. Yin SYL, Tserng HP, Wang JC, Tsai SC (2009) Developing a precast production management system using RFID technology. Autom Constr 18(5):677–691

13. Motamedi A, Hammad A (2009) Lifecycle management of facilities components using radio frequency identification and building information model. Electron J Inf Technol Constr 14(2009):238–262
14. Jaselskis EJ, Anderson MR, Jahren CT, Rodriguez Y, Njos S (1995) Radio frequency identification applications in construction industry. J Constr Eng Manag 121(2):189–196
15. Chen Z, Li H, Wong TC (2002) An application of bar-code system for reducing construction wastes. Autom Constr 11(5):521–533
16. Jaselskis EJ, El-Misalami T (2003) Implementing radio frequency identification in the construction process. J Constr Eng Manag 129(6):80–688
17. Oloufa AA, Ikeda M, Oda H (2003) Situational awareness of construction equipment using GPS, wireless and web technologies. Autom Constr 12(6):737–748
18. Jang WS, Skibniewski MJ (2007) Wireless sensor technologies for automated tracking and monitoring of construction materials utilizing Zigbee networks. In: ASCE Construction Research Congress: The Global Construction Community, Grand Bahamas Island
19. Song J, Haas CT, Caldas CH (2007) A proximity-based method for locating RFID tagged objects. Adv Eng Inform 21(4):367–376
20. Song J, Haas CT, Caldas CH (2006) Tracking the location of materials on construction job sites. J Constr Eng Manag 132(9):911–918
21. Caron F, Razavi SN, Song J, Vanheeghe P, Duflos E, Caldas CH, Haas CT (2007) Locating sensor nodes on construction projects. Auton Robot 22(3):255–263
22. Ergen E, Akinci B, Sacks R (2007) Life-cycle data management of engineered-to-order components using radio frequency identification. Adv Eng Inform 21(4):356–366
23. Ergen E, Akinci B, Sacks R (2007) Tracking and locating components in a precast storage yard utilizing radio frequency identification technology and GPS. Autom Constr 16(3):354–367
24. Yu SN, Lee SY, Han CS, Lee KY, Lee SH (2007) Development of the curtain wall installation robot: performance and efficiency tests at a construction site. Auton Robot 22(3):281–291
25. Tzeng CT, Chiang YC, Chiang CM, Lai CM (2008) Combination of radio frequency identification (RFID) and field verification tests of interior decorating materials. Autom Constr 18(1):16–23
26. Jang WS, Skibniewski MJ (2008) A wireless network system for automated tracking of construction materials on project sites. J Constr Eng Manag 14(1):11–19
27. Torrent DG, Caldas CH (2009) Methodology for automating the identification and localization of construction components on industrial projects. J Comput Civ Eng 23(1):3–13
28. Majrouhi Sardroud J, Limbachiya MC (2010) Improving construction supply chain management with integrated application of RFID technology and portal system. In: The 8th International Conference on Logistics Research (RIRL 2010), Sept. 29–30 and Oct. 1st, 2010, Bordeaux, France
29. Majrouhi Sardroud J, Limbachiya MC (2010) Integrated advance data storage technology for effective construction logistics management. In: 27th International Symposium on Automation and Robotics in Construction (ISARC 2010), June 25–27, 2010, Bratislava, Slovakia
30. Naresh AL, Jahren CT (1997) Communications and tracking for construction vehicles. J Constr Eng Manag 123(3):261–268
31. Sacks R, Navon R, Brodetskaia I, Shapira A (2005) Feasibility of automated monitoring of lifting equipment in support of project control. J Constr Eng Manag 131(5):604–614
32. Goodrum PM, McLaren MA, Durfee A (2006) The application of active radio frequency identification technology for tool tracking on construction job sites. Autom Constr 15(3):292–302
33. Lee UK, Kang KI, Kim GH, Cho HH (2006) Improving tower crane productivity using wireless technology. Computer-Aided Civ Inf Eng 21(8):594–604
34. Lu M, Chen W, Shen X, Lam HC, Liu J (2007) Positioning and tracking construction vehicles in highly dense urban areas and building construction sites. Autom Constr 16(5):647–656

35. Nambiar AN (2009) RFID Technology: A Review of its Applications. Lect Notes Eng Comput Sci 2179(1):1253–1259
36. Huang X (2008) Efficient and reliable estimation of tags in RFID systems. Lect Notes Eng Comput Sci 2169(1):1169–1173
37. Majrouhi Sardroud J, Limbachiya MC (2010) Utilization of advanced data storage technology to conduct construction industry on clear environment. In: International Conference on Energy, Environment, and Sustainable Development (ICEESD 2010), June 28–30, 2010, Paris, France
38. Landt J (2005) The history of RFID. IEEE Potentials 24(4):8–11
39. ERABUILD (2006) Review of the current state of Radio Frequency Identification (RFID) technology, its use and potential future use in construction. National Agency for Enterprise and Construction, Tekes, Formas and DTI, Final Report
40. Lahiri S (2005) RFID sourcebook. IBM Press, Upper Saddle River
41. Su CJ, Chou TC (2008) An radio frequency identification and enterprise resource planning-enabled mobile asset management information system. Lect Notes Eng Comput Sci 2169(1):1837–1842
42. Kaplan ED, Hegarty CJ (2006) Understanding GPS, principles and applications. Artech House, Inc., Norwood
43. Xu G (2007) GPS: theory algorithms and applications. Springer, Berlin
44. Kupper A (2005) Location-based services, fundamentals and operation. Wiley, West Sussex
45. French GT (1996) Understanding the GPS—an introduction to the global positioning system. GeoResearch, Inc., Bethesda
46. Peyret F, Betaille D, Hintzy G (2000) High-precision application of GPS in the field of real-time equipment positioning. Autom Constr 9(3):299–314
47. Ward M, Thorpe T, Price A, Wren C (2004) Implementation and control of wireless data collection on construction sites. Electron J Inf Technol Constr (ITcon) 9:297–311

Chapter 18
A Case Study Analysis of an E-Business Security Negotiations Support Tool

Jason R. C. Nurse and Jane E. Sinclair

Abstract Active collaboration is undoubtedly one of the most important aspects within e-business. In addition to companies collaborating on ways to increase productivity and cut costs, there is a growing need for in-depth discussion and negotiations on their individual and collective security. This paper extends previous work on a tool aimed at supporting the cross-enterprise security negotiations process. Specifically, our goal in this article is to briefly present a case study analysis and evaluation of the usage of the tool. This provides further real-world insight into the practicality of the tool and the solution model which it embodies.

18.1 Introduction

E-business has matured into one of the most cost-efficient and streamlined ways of conducting business. As the use of this new business paradigm thrives however, ensuring adequate levels of security for these service offerings emerges as a critical goal. The need for security is driven by an increasing regulatory and standards requirements base (e.g. EU Data Protection Act and US Sarbanes–Oxley Act) and escalating security threats worldwide (as indicated in [1]). Similar to the business-level collaborations necessary to facilitate these interactions, there also needs to be a number of discussions and negotiations on security. A key problem during collaborations however is the complex discussion task that often ensues as

J. R. C. Nurse (✉) · J. E. Sinclair
Department of Computer Science, Warwick University, Coventry, CV4 7AL, UK
e-mail: jnurse@dcs.warwick.ac.uk

J. E. Sinclair
e-mail: jane.sinclair@dcs.warwick.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_18,
© Springer Science+Business Media B.V. 2011

companies have different security postures, a range of disparate security needs, may have dissimilar laws/regulations which they each subscribe to, have different skill sets/experience levels and so on. Owing to these and other challenges, [2] aptly labels the related process as 'security mayhem'.

With appreciation of the collaboration difficulties highlighted above, particularly in terms of security approaches in Web services-based interactions, in previous work we have presented BOF4WSS, a Business-Oriented Framework for enhancing Web Services Security for e-business [3, 4]. The framework's novelty stemmed from its concentration on a cross-enterprise development methodology to aid collaborating e-businesses in jointly creating secure and trusted interactions. Additionally, BOF4WSS aims to fit together a majority of the critical pieces of the WS security puzzle (for example, key new approaches such as [5, 6]) to propose a well-rounded, highly structured, extensible framework (framework and methodology being synonymous in the context of this work).

Progressing from the BOF4WSS methodology itself, our emphasis has shifted to supplying software to support it and assist in its seamless application to business scenarios. In previous articles we have presented (see [7]) and initially evaluated (see [8]) one of these tools, which was developed to support and ease security negotiations across collaborating e-businesses. In terms of BOF4WSS, this refers specifically to easing the transition from the individual Requirements Elicitation stage to the subsequent joint Negotiations stage. Generally, some of the main problems identified and targeted included, understanding other companies' security documentation, understanding the motivation behind partnering businesses' security needs/decisions, and being able to easily match and compare security decisions from entities which target the same situation and risk. Related work in [9, 10] and feedback from interviewed security practitioners supports these issues.

Building on previous research, the aim of this paper therefore is to extend initial evaluation work in [8] and pull together the compatibility evaluation of the tool and the Solution model it embodies through the use of a case scenario analysis. This enables a more complete evaluation of the proposals because, unlike the compatibility assessment in [8], it progresses from the initial model stages to the final tool output produced. The scenario contains two companies using two popular system-supported security Risk Management/Assessment (RM/RA) methods. Topics to be considered in this analysis include: how tool data is transferred to the RM/RA approaches/software (as expected in the Solution model [7]); How is typical RM/RA approach information represented in the tool's common, custom-build XML-based language; and finally, how close, if at all, can the tool bring together the different RM/RA approaches used by companies to ease stage transition within BOF4WSS. If the tool can interplay with a majority of the security-related information output from popular RM/RA techniques, its feasibility as a system that can work alongside current approaches used in businesses today, will be evidenced.

The next section of this paper reviews the Solution model and resulting tool to support security negotiations across e-businesses. Then, in Sect. 18.3, we give a brief background on the business scenario and begin the case study analysis.

Findings are discussed as they are found. Section 18.4 completes this contribution by providing conclusions and outlining directions for future work.

18.2 Solution Model and Tool: A Recap

The Solution model is the conceptual base for the software tool developed in our research. It was initially presented in [7] and consists of four component stages. These are: Security Actions Analysis, Ontology Design, Language Definition and Risk Catalogue Creation.

Security actions analysis This stage focuses on reviewing the literature in the security risk management field, and critically examining how security actions and requirements are determined. A *security action* is broadly defined as any way in which a business handles the risk it faces (e.g. 'insurance will be purchased to protect against very skilled and sophisticated hacker attacks'), and a *security requirement* is a high-to-medium level desire, expressed to mitigate a risk (e.g. 'classified information must be encrypted when transferred over a network connection'). The key outcome of this stage is a thorough understanding of the relevant security domain which could then be used as a foundation for future stages.

Ontology design The aim of this component is to produce a high-level ontology design using the findings from the previous stage, to establish a common understanding and semantics structure of the security actions (and generally security risk management) domain. This common or shared understanding is a critical prerequisite when considering the difficulties businesses face (because of different terminologies used, RM/RA methods applied, and so on) as they try to understand their partners' security documentation which is supplied in BOF4WSS' Negotiations phase. Further detail on the Security Actions Analysis and Ontology Design stages (inclusive of draft ontology) can be seen in [11].

Language definition This stage has two sub-components. First is the development of a XML-based language called Security Action Definition Markup Language (SADML). This allows for the establishment of a common format (based on the ontology) by which security actions/requirements information provided by companies is formally expressed, and also later processed by the resulting tool. Second is the proposal of a user-friendly interface such as a data entry screen or template document by which businesses' security-related data could be entered, and subsequently marked up in SADML. This interface would act as a guide for companies in prompting them to supply complete information as they prepare to come together for negotiations.

Risk catalogue creation This final component stage addresses the problem of matching and comparing security actions/requirements across enterprises by defining a shared risks catalogue. Given that businesses use risks from this shared catalogue as input to their RM/RA methods, regardless of the security actions that they decide individually, the underlying risks could be used by the tool to automatically match their actions. To increase flexibility, the catalogue would feature an extensive and updatable set of security risks.

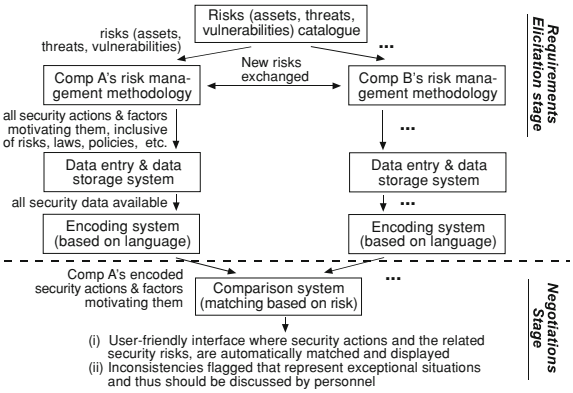

Risks (assets, threats, vulnerabilities) catalogue
risks (assets, threats, vulnerabilities)
New risks exchanged
Comp A's risk management methodology
Comp B's risk management methodology
all security actions & factors motivating them, inclusive of risks, laws, policies, etc.
Data entry & data storage system
Data entry & data storage system
all security data available
Encoding system (based on language)
Encoding system (based on language)
Comp A's encoded security actions & factors motivating them
Comparison system (matching based on risk)
(i) User-friendly interface where security actions and the related security risks, are automatically matched and displayed
(ii) Inconsistencies flagged that represent exceptional situations and thus should be discussed by personnel
Requirements Elicitation stage
Negotiations Stage

Fig. 18.1 Process flow of implemented solution model

With a recap of the Solution model now provided, Fig. 18.1 shows a process flow of how the implemented model i.e. the tool, works. In this diagram, Comp A and Comp B are companies using BOF4WSS for an online business scenario.

To explain the process flow: First, companies would select a set of risks from the catalogue that apply to their particular business scenario, and use these as input to their different RM/RA methodologies. Any new risks to be considered which are not available in the catalogue can be exchanged for this scenario. After companies have used their RM/RA approaches to determine their individual security actions (inclusive of motivational factors), these are then input into the Data entry and storage system. This system uses a user-friendly interface to read in the data (as suggested in the Language Definition stage), and stores it to a back-end database to allow for data retrieval, updating and so on. This interface, and generally the tool, mirror the understanding of concepts defined in the ontology.

As companies are about to come together for Negotiations, the Encoding system is used to read security data from the database and encode it into SADML. In the Negotiations stage of BOF4WSS, companies bring their individual SADML documents and these are passed to the tool's Comparison system. This system matches companies' security actions based on risks which they address, and aims to provide a user-friendly interface in which (i) security actions can be quickly compared and discussed, (ii) any inconsistencies would be flagged for follow-up by personnel, and (iii) a shared understanding of security terms, risks and so on, will be upheld due to the references that can be made to the ontology. Next, in Sect. 18.3, we conduct the case study analysis to give further insight into the use and practicality of the tool proposed.

18.3 Case Study Analysis

The core aim of this section is to complete the compatibility and feasibility evaluation first presented in [8], using a full case study analysis. In that previous work, a very detailed discourse and a number of mappings were presented. Now

the objective is to put that and other aspects of the Solution model and tool into a more real-world context. In addition to further supporting the feasibility of this research's proposals, this would enable a more thorough evaluation of the model as it progressed from the initial Central Risk Catalogue to the final tool output.

The case scenario to be used consists of two businesses, Buyer and Supplier. These companies have worked with each other in the past using mainly manual and other offline interactions. To enable their processes to be more integrated and streamlined, the parties are now choosing to use the Internet and WS technology suite for online business-to-business communications. As security is a key priority for companies, they are adopting BOF4WSS to aid in the creation of a secure WS-based business scenario. In line with this paper, the areas of focus are the progression from the Requirements Elicitation stage to the Negotiation stage. This involves the passing and then negotiation on entities' security needs and requirements.

In terms of RM/RA and determining security needs and requirements, EBIOS and CORAS are the two methods used by entities. EBIOS is a risk management approach for assessing and treating risks in the field of information systems security [12]. CORAS is a tool-supported methodology for model-based risk assessment of security-critical systems [13]. Specifically, to analyze risk and determine security actions, Buyer uses EBIOS and its software, whereas Supplier employs CORAS and its supporting tool. Next, we begin the case study analysis.

According to the Solution model flow (see Fig. 18.1), regardless of the RM/RA method used, the starting point of the scenario should be a common risks base or catalogue. This point however is where one of the first difficulties in the evaluation surfaced. When the model was first conceived it was assumed that the transferring of common risk data to RM/RA approaches would be done manually. During the completion of this study however, such a process actually proved somewhat tedious. This is especially in terms of accurate and consistent mapping of data from the common risks catalogue to the RM/RA methods and software.

If there was a risk to the confidentiality of Web services messages in the Risk Catalogue system therefore, the problem was: how could that data and the related data on vulnerabilities, threats and assets, be quickly, accurately and consistently entered into the RM/RA approaches and their software. Figure 18.2 depicts the area of focus in the 'Process flow of the implemented Solution model' diagram (Fig. 18.1).

Possibly the best solution to this problem resides in the automated mapping of data from the Central Risk Catalogue to the RM/RA method software, which in this case is represented by the EBIOS and CORAS tools (used by Buyer and Supplier respectively). Two options were identified by which this could occur. The first option consisted of adding an export capability to the Central Risk Catalogue system, which would output data on risks in the machine-readable formats of common RM/RA approach software. This is beneficial because it would be a central point where numerous RM/RA software formats could be generated. Furthermore, it could take advantage of the 'Import' and 'Open File/Project', functionalities which are standard in a number of RM/RA software. For example, both CORAS and EBIOS tools have these capabilities.

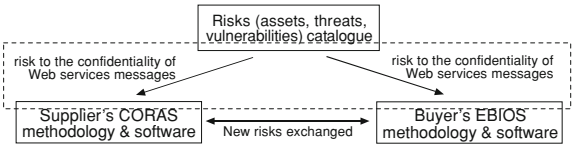

Risks (assets, threats, vulnerabilities) catalogue
risk to the confidentiality of Web services messages
risk to the confidentiality of Web services messages
Supplier's CORAS methodology & software
New risks exchanged
Buyer's EBIOS methodology & software

Fig. 18.2 Area of focus in process flow of implemented solution model

One caveat noticed when assessing the Risk Catalogue export capability option is that unique identification numbers (IDs) for elements (e.g., *Menace* IDs in EBIOS or *risk-analysis-result* IDs in CORAS) generated by the Central system might conflict with the same element IDs generated by the actual software running at each company. There would therefore need to be some agreed allotment of ID ranges for the Catalogue-based option to function properly.

The second option suggests a more decentralized implementation where extensions could be added to the RM/RA software systems to enable them to read in and process Risk Catalogue system data. This would avoid the problem of conflicting IDs, but introduces the need to access, understand and edit various software systems. For this case, EBIOS and CORAS are good candidates in this regard as both are open source implementations (see [12] and [13] respectively).

Apart from the programming that would be necessary in both options above, there is the question of exactly how to map Risk Catalogue system data to EBIOS and CORAS. This however can be largely addressed by reversing the mapping tables used as the basis for previous evaluation work in [8]. This is because the tool's Entity Relationship Diagram (ERD) is not dissimilar to that of the Risk Catalogue system. Essentially, one would now be going from database records to EBIOS and CORAS software XML formats. *Risk*, *ProjectRisk*, *Asset*, *Vulnerability* and *Threat* are some of the main database tables mapped in [8] that would be used in reverse to map risks data from the Catalogue system.

Having briefly digressed from the case study to discuss how transferring data from the shared risks catalogue could be addressed, the focus resumes at the RM/RA software stage. This relates to the bottom two boxes in Fig. 18.2. After Supplier and Buyer have agreed the risks to be used, they conduct their individual analyses. This generally encompasses the processes of risk estimation, risk evaluation and treatment. The two code snippets below give an initial idea of the data generated by each entity's RM/RA method. This and most of the following examples are based around a security risk defined by companies relating to the integrity and confidentiality of Web services messages passed between them during online interactions. Hereafter, this is referred to simply as Risk101; 'Risk101' is also used as the lower-level ID value originating from the risks catalogue which is employed in each company's RM/RA software. From the code snippets, one can see exactly how different the representations of the same risk may be from company to company. As would be expected, a similar reality exists regarding the other types of data produced (e.g. related to risk factors, risk estimates, security actions and so on). The + sign in the code indicates that there is additional data which is not displayed/expanded considering space limitations.

```
<Risk ID="RiskScenario.1252746098288" label="Risk101"
      menace="Menace.1050437920519"
      description="The integrity and confidentiality of data in a Web services' message
      (in transit) is compromised" sof="AttackPotential.1070307963407" coverLevel=
      "SecurityObjectiveCoverValue.1078561424090" ... >
+ <ScenarioPotentiality potentiality="Potentiality.1076645892186">
</Risk>
```

Code Snippet #-1. EBIOS (Buyer) representation of the risk

```
<row>
 <cell columnId="riskId">Risk101</cell>
 <cell columnId="assetId">WSMessage</cell>
 <cell columnId="incident">Eavesdropping and tampering with data in a Web services'
   message (in transit)</cell>
 <cell columnId="consequenceValue">Medium</cell>
 <cell columnId="frequencyValue">Low</cell>
 <cell columnId="scenario"/>
</row>
```

Code Snippet #-2. CORAS (Supplier) representation of the risk

With the RM/RA methodologies at each business complete, the next step was mapping the output data from Buyer and Supplier to the tool. This process was covered in detail in [8] and therefore is not analyzed in depth here. From a case study perspective however, one intriguing additional observation was made—that is, although RM/RA methods did not accommodate certain data in a very structured way as expected by the tool, it did not mean that the data was not present in companies' considerations.

In Supplier's CORAS software output shown in Code Snippet 3 for example, it is apparent that a limited security budget influenced Supplier's treatment strategy decision (see *treatmentDescription columnId*). Any automated mapping to our tool therefore should ideally capture this data as a unique Risk Treatment factor. To recap, a treatment factor is an aspect that influences or in some way motivates a particular treatment for a risk. Common examples are laws, regulations, security policies, limited budgets and contractual obligations. Capturing this treatment data was not possible however because the machine-readable output of CORAS does not distinctly define such aspects in its XML structure. Here it is just in plain text.

```
<row>
 <cell columnId="treatmentId">TRT101</cell>
 <cell columnId="riskOrCategoryId">Risk101</cell>
 <cell columnId="treatmentStrategy">Retain</cell>
 <cell columnId="treatmentDescription">The unlikeliness of this risk and a limited
   security budget are the reasons for risk acceptance</cell>
 <cell columnId="treatmentReferences">Threat_Analysis09.doc</cell>
</row>
```

Code Snippet #-3. CORAS representation of a risk treatment

A similar situation is present in Buyer's EBIOS output regarding risk estimation. In this case, Buyer has used EBIOS to prioritize risks, however, because their technique is so elaborate it does not allow for a clear and reliable automated mapping to the risk level concepts in the tool.

To tackle these mapping issues a few other techniques were assessed but manual mapping proved to be the only dependable solution. This mapping involved noting the type of data requested by the tool (such as influential security policies or budgetary limitations) and using its data entry screens to manually enter that data. This was easily done in this case through the creation of a *TreatmentFactor* record in the tool and then linking that record to the respective risk treatment, formally the *SecurityAction* database record. The *TreatmentFactor* table is used to store elements that influence or affect the treatment of risks. Examples of such were mentioned previously.

Regarding the manual risk estimation and prioritization mapping needed for EBIOS mapping, a level of subjectivity would be introduced as users seek to map values in their analysis to the risk levels expected in the tool. To compensate for this subjectivity, detailed justifications and descriptions of chosen risk levels should be provided by parties. This information would be entered in the tool's respective *RiskEstimate* database record's *risk_level_remarks*, *probability_remarks*, *impact_remarks* and *adequacy_of_controls_remarks* fields. (The *RiskEstimate* table defines the value of a risk, the probability and impact of it occurring, and the effectiveness of current controls in preventing that risk.) Generally, at the end of mapping, companies' personnel should browse screens in the tool to ensure that all the required information has been transferred.

The next step in the case study was encoding each business's mapped data (now in the tool's database) to SADML documents. This process went without error. In Code Snippet 4, an example of the security risk under examination (Risk101) is presented. The marked-up risk data has the same basis across businesses and documents due to the use of the shared risks base in the beginning. SADML provides the common structure, elements and attribute names. Different companies may add varying comments or descriptions however. The specific code in Snippet 4 is from Buyer.

```
<risk id="RISK101"><threats>
 <threat>
  <name>Eavesdropping and tampering with data in a Web services' message
      (in transit)</name>
  <threatAgent><agentName>Malicious party</agentName><comment/>
  </threatAgent>
  <comment />
 </threat></threats><vulnerabilities>
 <vulnerability>
  <name>Circulating information in inappropriately secured formats</name>
  <asset>
   <dtype>property:data</dtype>
   <assetName>web service message</assetName>
   <comment>The data carried in the message is the key aspect</comment>
  </asset>
  <comment />
 </vulnerability></vulnerabilities>
 <riskComment>Violation of confidentiality using eavesdropping</riskComment>
 ...
</risk>
```

Code Snippet #-4. SADML representation of the highlighted risk

The real difference in SADML documents across Buyer and Supplier is visible when it comes to the treatment of Risk101. In this case, Buyer aims to mitigate this risk while Supplier accepts it. SADML Code Snippet 5 shows this and the respective treatment factors. On the left hand side is Buyer's document and on the right, Supplier's. The + sign indicates that there is additional data which is not displayed here.

```
<mitigationAction>
  <name>Protect against eavesdropping
    on Web service messages being
    transmitted between partners</name>
  <details>The organization must take
    measures to ensure there is no
    eavesdropping on data being
    transmitted between Web services
    across business parties.</details>
  <risks>
+   <risk id="RISK101">
  </risks>
+ <lawAndRegulationRefs>
  <contractualObligationRefs />
  <businessPolicyRefs />
+ <securityPolicyRefs>
  <securityBudgetRefs />
+ <securityRequirementRefs>
</mitigationAction>
```

```
<acceptanceAction>
  <name>The unlikeliness of this risk and
    a limited security budget are the
    reasons for risk acceptance</name>
  <details>Threat_Analysis09.doc</details>
  <risks>
+   <risk id="RISK101">
  </risks>
  <lawAndRegulationRefs />
  <contractualObligationRefs />
  <businessPolicyRefs />
  <securityPolicyRefs>
+ <securityBudgetRefs>
+ <riskActionImplementationDetailRefs />
</acceptanceAction>
```

Code Snippet #-5. SADML representations of companiesGúø risk treatment choices

When compared to the original output from EBIOS and CORAS, one can appreciate the use of the standard format supplied by SADML. In this respect, SADML provides a bridge between different RM/RA methods and their software systems, which can then be used as a platform to compare high-level security actions across enterprises. It is worth noting that the benefits possible with SADML are largely due to its foundation in the well-researched ontology from the Solution model [7, 11].

With all the preparatory stages in the case process completed, Fig. 18.3 displays the output of the model's final stage i.e. the tool's Comparison System which is presented to personnel at Buyer and Supplier. Apart from the user-friendly, colour-coded report, the real benefit associated with this output is the automation of several of the preceding steps taken to reach this point. These included: (i) gathering data from RM/RA approaches (such as EBIOS and CORAS), albeit in a semi-automated fashion; (ii) allowing for influential factors in risk treatment that are key to forthcoming negotiations, to be defined in initial stages; and finally, (iii) matching and comparing the security actions and requirements of companies based on shared risks faced.

The output in Fig. 18.3 also aids in reconciling semantic differences across RM/RA approaches as these issues are resolved by mapping rules earlier in the process (as covered in [8]). Furthermore, personnel from companies can refer to the

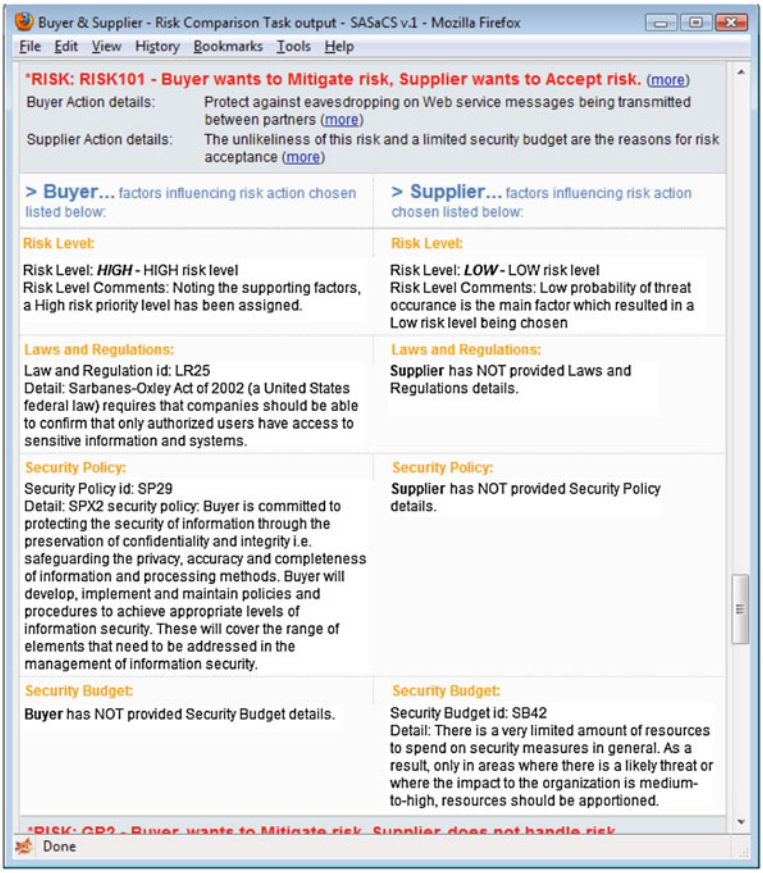

Buyer & Supplier - Risk Comparison Task output - SASaCS v.1 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
*RISK: RISK101 - Buyer wants to Mitigate risk, Supplier wants to Accept risk. (more)
Buyer Action details: Protect against eavesdropping on Web service messages being transmitted between partners (more)
Supplier Action details: The unlikeliness of this risk and a limited security budget are the reasons for risk acceptance (more)
> Buyer... factors influencing risk action chosen listed below:
> Supplier... factors influencing risk action chosen listed below:
Risk Level:
Risk Level: HIGH - HIGH risk level
Risk Level Comments: Noting the supporting factors, a High risk priority level has been assigned.
Risk Level:
Risk Level: LOW - LOW risk level
Risk Level Comments: Low probability of threat occurance is the main factor which resulted in a Low risk level being chosen
Laws and Regulations:
Law and Regulation id: LR25
Detail: Sarbanes-Oxley Act of 2002 (a United States federal law) requires that companies should be able to confirm that only authorized users have access to sensitive information and systems.
Laws and Regulations:
Supplier has NOT provided Laws and Regulations details.
Security Policy:
Security Policy id: SP29
Detail: SPX2 security policy: Buyer is committed to protecting the security of information through the preservation of confidentiality and integrity i.e. safeguarding the privacy, accuracy and completeness of information and processing methods. Buyer will develop, implement and maintain policies and procedures to achieve appropriate levels of information security. These will cover the range of elements that need to be addressed in the management of information security.
Security Policy:
Supplier has NOT provided Security Policy details.
Security Budget:
Buyer has NOT provided Security Budget details.
Security Budget:
Security Budget id: SB42
Detail: There is a very limited amount of resources to spend on security measures in general. As a result, only in areas where there is a likely threat or where the impact to the organization is medium-to-high, resources should be apportioned.
Done

Fig. 18.3 Area of focus in process flow of implemented solution model

ontology and the inclusive shared definitions/terminology at any point. This would be done to attain a clear understanding on terms in the context of the interactions. As parties come together therefore, they can immediately identify any conflicts in treatment choices and have the main factors supporting those conflicting choices displayed. This and the discussion above give evidence to show that in many ways, our tool has brought the interacting enterprises closer together—particularly in bridging a number of key gaps across companies. This therefore allows for an easier transition between the Requirement Elicitation and Negotiation phases in BOF4WSS.

The shortcomings of the tool identified in this section's case study centered on the manual effort needed at a few stages to complete data mapping. This acted to limit some of the Solution model's automated negotiations support goals. To critically consider this point however, the level of automation and support that is present now would significantly bridge the disparity gaps and support a much

easier negotiation on security actions between parties. A small degree manual intervention therefore, even though not preferred, might be negligible. This is especially in business scenarios where there are large amounts of risks or security actions to be deliberated, and thus saving time at any point would result in substantial boosts in productivity.

18.4 Conclusion and Future Work

The main goal of this paper was to extend initial evaluation work in [8] and pull together the compatibility evaluation of tool and generally the Solution model it embodies through the use of a full case study analysis. The findings from this new and more complete analysis are seen to supply further evidence to support the tool as a useful, feasible and practical system to aid in cross-enterprise security negotiations. This is especially in terms of BOF4WSS but there might also be other opportunities for its use in other collaborative e-business development methodologies. The main benefit of the tool and model are to be found in a much easier negotiation process which then results in significantly increased productivity for companies.

There are two prime avenues for future work. The first avenue consists of testing the tool with other RM/RA techniques; IT-Grundschutz Manual [14] and NIST Risk Management Guide for Information Technology Systems SP800-30 [15] are some of the methods under investigation for this task. Positive evaluation results would further support the tool and any justified nuances of those popular techniques would aid in its refinement. The second avenue is more generic and looks towards the research and development of additional approaches and systems to support BOF4WSS. Considering the comprehensive and detailed nature of the framework, support tools could be invaluable in promoting BOF4WSS's use and seamless application to scenarios.

References

1. PricewaterhouseCoopers LLP. Information Security Breaches Survey 2010 [Online]. Available: http://www.pwc.co.uk/eng/publications/isbs_survey_2010.html
2. Tiller JS (2005) The ethical hack: a framework for business value penetration testing. Auerbach Publications, Boca Raton
3. Nurse JRC, Sinclair JE (2009) BOF4WSS: a business-oriented framework for enhancing web services security for e-Business. In: 4th International Conference on Internet and Web Applications and Services. IEEE Computer Society, pp 286–291
4. Nurse JRC, Sinclair JE (2009) Securing e-Businesses that use Web Services — A Guided Tour through BOF4WSS. Int J Adv Internet Technol 2(4):253–276
5. Steel C, Nagappan R, Lai R (2005) Core security patterns: best practices and strategies for J2EE™, web services and identity management. Prentice Hall PTR, Upper Saddle River
6. Gutierrez C, Fernandez-Medina E, Piattini M (2006) PWSSec: process for web services security. In: IEEE International Conference on Web Services, pp 213–222

7. Nurse JRC, Sinclair JE (2010) A solution model and tool for supporting the negotiation of security decisions in e-business collaborations. In: 5th International Conference on Internet and Web Applications and Services. IEEE Computer Society, pp 13–18
8. Nurse JRC, Sinclair JE (2010) Evaluating the compatibility of a tool to support e-businesses' security negotiations. In: Lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering 2010, WCE 2010, London, UK, pp 438–443
9. Yau SS, Chen Z (2006) A framework for specifying and managing security requirements in collaborative systems. In: Yang LT, Jin H, Ma J, Ungerer T (eds) Autonomic and Trusted Computing, ser. Lecture Notes in Computer Science, vol 4158. Springer, Heidelberg, pp 500–510
10. Todd M, Zibert E, Midwinter T (2006) Security risk management in the BT HP alliance. BT Technol J 24(4):47–52
11. Nurse JRC, Sinclair JE (2009) Supporting the comparison of business-level security requirements within cross-enterprise service development. In: Abramowicz W (ed) Business Information Systems, ser. Lecture Notes in Business Information Processing, vol 21. Springer, Heidelberg, pp 61–72
12. DCSSI (2004) Expression des besoins et identification des objectifs de securite (EBIOS)—Section 1–5, Secretariat General de la Defense Nationale. Direction Centrale de la Securitec des Systecmes D'Information, Technical Report
13. den Braber F, Braendeland G, Dahl HEI, Engan I, Hogganvik I, Lund MS, Solhaug B, Stolen K, Vraalsen F (2006) The CORAS model-based method for security risk analysis. SINTEF, Technical Report
14. Federal Office for Information Security (BSI). IT-Grundschutz Manual [Online]. Available: https://www.bsi.bund.de/EN/Topics/ITGrundschutz/itgrundschutz_node.html
15. National Institute of Standards and Technology (NIST) (2002) Risk management guide for information technology systems (Special Publication 800-30), Technical Report

Chapter 19
Smart Card Web Server

Lazaros Kyrillidis, Keith Mayes and Konstantinos Markantonakis

Abstract In this article (based on "Kyrillidis L, Mayes K, Markantonakis K (2010) Web server on a SIM card. Lecture notes in engineering and computer science: Proceedings of the World Congress on Engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, pp 253–259") we discuss about the integration of a web server on a SIM card and we attempt an analysis from various perspectives (management, operation, security). A brief representation of the Smart Card Web Server (SCWS) will take place along with a use case that will help the reader to identify the way that an SCWS can be used in practice, before we reach to a final conclusion.

19.1 Introduction

The World Wide Web (WWW) was a major step forward for humanity in terms of communication, information and entertainment. Originally, the web pages were static, not being changed very often and without any user interaction. This lack of interactivity led to the creation of server side scripting languages (like PHP) that allowed the creation of dynamic pages. These pages are often updated according to the users' interests and in recent years even their content is created from the users (blogs, social networking, etc.). In order for these pages to be properly created

L. Kyrillidis (✉)
Information Security Strategy Consultant, Agias Lavras 3, Neapoli, Thessaloniki, Greece
e-mail: lazaroskyr4@yahoo.gr

K. Mayes · K. Markantonakis
Smart Card Centre, Royal Holloway, University of London, London, UK
e-mail: keith.mayes@rhul.ac.uk

K. Markantonakis
e-mail: K.Markantonakis@rhul.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*,
Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_19,
© Springer Science+Business Media B.V. 2011

and served, a special type of computer program is required. This program known as a Web Server accepts the users' requests, processes them and returns the result to the requesting user's browser.

Another important step for modern communications was the invention of mobile phones. The first devices suffered from fraud, because it was quite easy to intercept communications and clone phones. This inevitably led to the introduction of a secure, tamper-resistant module that could be used for securing the storage of sensitive information and cryptographic keys/algorithms that were used to secure communication between the phone and the network provider's facilities. This module is referred to as the Subscriber Identity Module (SIM).

The idea of hosting a web server on a SIM was proposed almost a decade ago [1] and although is not yet commercially available, recent technological advances suggest that the idea could be reconsidered. While the integration of SIM and web server could offer new fascinating prospects to both the network providers and the users, there is a security concern that it might also help attackers to gain access to the SIM contents. A practical concern is the extent of the added management, operation and personalization costs that this integration would entail.

19.2 Web Server on A Sim Card

The Open Mobile Alliance is an international body that is formed to produce and promote standards for the mobile communications industry in order to encourage the interoperability of products aiming at lower operational costs and higher quality products for the end users [2].

One of the standards that OMA created, was the web server on the SIM card standard (Smart Card Web Server—SCWS) [3–5]. This standard defines a number of entities that the web server must contain:

- *SCWS* The web server itself. It is located inside the SIM card.
- *SCWS gateway* This entity is needed when the SIM card cannot directly respond to HyperText Transport Protocol (HTTP) requests, so the gateway's main purpose is to translate the browser's requests from HTTP to the local transport protocol and vice versa. A common local protocol would be the Bearer Independent Protocol (BIP) [6]. Additionally, the gateway is proposed to host a form of Access Control List (ACL) to control the access to the SCWS. It is located on the phone.
- *HTTP (s) Client* The browser that will initiate requests towards the SCWS and will present the response to the end user. It is located on the phone.
- *SCWS Administration Application* This entity is used for SCWS software updates/patches that would be applied remotely to the SCWS and for installing and/or updating possible web applications that may run on the SIM card. Additionally, it may be used to send new content in the form of HTML pages to the SCWS. It is located in the network provider's premises.

19.2.1 Communication Protocols

The SCWS will use two different protocols to communicate with entities outside of the SIM card. The first will be the BIP protocol to encapsulate HTTP(s) packets when the SIM card does not implement its own TCP/IP stack, while the second one will be the HTTP(s) for the newer smart cards that will allow direct HTTP access. We will now take a more detailed look at these two protocols and how they can be used with SCWS (according to [3−5]):

19.2.1.1 BIP Protocol

As mentioned earlier, the BIP protocol will allow incoming and outgoing communication when the SIM card cannot support direct HTTP access. The SIM card will work into two modes:

- *Client mode* The SCWS communicates with the remote administration entity to receive updates. The gateway translates requests from BIP to HTTP(s) (the SIM card can "speak" BIP, while the remote administration will "speak" HTTP(s)).
- *Server mode* The SCWS communicates with the browser. The gateway is once again present, executing the translation between the BIP protocol that the SIM card understands and the HTTP(s) requests/answers that the browser understands.

19.2.1.2 TCP/IP Protocol

If the SIM card implements its own TCP/IP stack (from Java card 3.0 and onwards), there will be no need for the gateway and the communication will take place directly between the two entities (either the SCWS and remote server or the SCWS and the browser).

19.2.2 Administration Protocols

When the SCWS is in client mode this means that it exchanges messages with the remote server. There are two ways for this message exchange to take place:

- When the amount of data that must be exchanged is relatively small, then the Lightweight Administration Protocol should be used. The bearer of the commands that encapsulate the data is an SMS or multiple SMS and when the SCWS receives these SMS(s), it parses it/them and then it must send a response back to the remote server, so that the latter can determine if it can send the next command or simply terminate the connection.
- A second way that can be used for administration purposes is the Full Administration Protocol. A card administration agent that is located inside the SIM card is responsible to encapsulate and transfer the HTTP messages over PSK-TLS,

to establish connection and if necessary reconnect if the connection is dropped. The agent sends a message to the remote server and the later responds with the administration command encapsulated within an HTTP response. The agent receives the command, passes it to SCWS which executes it. When the command is executed, the agent contacts the remote server for the next command. This operation continues until the remote server terminates the connection.

19.2.3 SCWS URL, IP, Port Numbers

As mentioned earlier, there will be two ways to communicate with the SCWS: either over HTTP or BIP. The port numbers that will be used when the HTTP requests are encapsulated inside BIP packets are 3516 for HTTP and 4116 for HTTPs. The format of the URL in both cases will be:

http://127.0.0.1:3516/file1/file2/test.html
https://127.0.0.1:4116/sec_file1/sec_test.html

If the access to the SCWS is provided directly over HTTP (s) the port numbers are the ones used for traditional web servers (80 for HTTP and 443 for HTTPs). The SIM card will now have its own IP address, so the loopback will no longer be needed. The format of the URL will be:

http://<smart_card_IP>[:80]/file1/file2/test.html
https://smart_card_IP>[:443]/sec_file1/sec_test.html

19.3 Using the SCWS for E-Voting

A possible use for the SCWS is given in the following example:

A country X provides to all its citizens an ID card that store (in addition to the citizen's name and ID card number), two certificates (one for encryption/decryption, one for digital signatures), the corresponding private keys and the government's public keys (for a similar example, see [7]). These certificates are also installed in a central location that is being administered by the government. The citizen can use his ID card for every transaction, either with the state or with other citizens. Additionally, the country has arranged with the mobile network providers to install these certificates/keys on the citizen's mobile phone, so that the later can use it to vote. The voting process is described in the following use case:

19.3.1 Process Flow

Let us suppose the following: $Cert_{A1}$ is the user's certificate, P_{A1} the private key and PU_{A1} the public key used for encryption/decryption and $Cert_{A2}$ is the user's

certificate, P_{A2} the private key and PU_{A2} the public key that are used for digital signatures. Likewise, $Cert_{B1}$ is the government's certificate, P_{B1} the private key and PU_{B1} the public key used for encryption/decryption and $Cert_{B2}$ is the government's certificate, P_{B2} the private key and PU_{B2} the public key that are used for digital signatures. Also let H be the hash algorithm that the two parties will use. The e-voting process is as follows:

- The network provider has updated the user's SCWS slightly by presenting a link on the user's home page named "e-voting".
- The user clicks on the link.
- The user's name and ID card number are encrypted with PU_{B1}; additionally both of them are hashed to produce the hash H_A and signed with P_{A2}. All these are sent to the government's remote server that hosts the e-voting site.

$$(ID_A,\ Name_A)_{PUB1}(H\ (ID_A,\ Name_A))_{PA2} \longrightarrow \text{Voting Server}$$

- The remote server decrypts $(ID_A, Name_A)_{PUB1}$ using P_{B1}, extracts PU_{A2} from $Cert_{A2}$ (which it already knows), verifies $(H\ (ID_A, Name_A))_{PA2}$ using PU_{A2} (and gets H_A), hashes the $(ID_A, Name_A)$ using H (and gets H_B), and checks H_A against H_B. If the two hashes much each other, the server authenticates the user and may proceed with the rest of the process. Then, it checks if the citizen has voted again and if not it creates a temporary entry in a database to show that the user's voting is in progress.
- After the user is authenticated to the server, he is presented with a link that points to the IP of the SCWS. The user clicks on the link and he is transferred to the SCWS environment. At the same time the remote server hashes the $(ID_A, Name_A)$, sends it signed with P_{B2} and also sends an encrypted link L which has embedded authentication data that will be used later on from the user (to authenticate himself on the remote site instead of providing a username/password):

$$(H(ID_A,\ Name_A))_{PB2}(L)_{PUA1} \longleftarrow \text{Voting Server}$$

- The SCWS receives the $(H\ (ID_A, Name_A))_{PB2}$ and verifies it using PU_{B2}. Then it hashes the user's ID and name that are stored on the SIM card with H and if the two hashes much each other, the server is authenticated and the SCWS can now prompt the user to provide the PIN. Additionally, the SCWS decrypts L using P_{A1}.
- The user provides the PIN, it is checked by the SCWS and if it is correct the SCWS displays the link L that points to the remote server.
- The user clicks on the link and can now browse the voting site and vote. When his voting is done, the permanent entry in the database is updated, to show that the user has voted.

19.3.2 Comments on this Use Case

Someone can argue about the need to use an SCWS for e-voting. While this document cannot explore the law and ethical issues that arise because of the

sensitive nature of the elections, there are some reasons that can justify its use.

The first is that a large part of the population is familiar with using a browser by using it in its day-to-day internet access. However, it is fairly easy for people to learn how to use it, even if they do not have previous experience.

Another important reason is that the security needed for the e-voting (and other similar uses like e-shopping) can be provided by using the SCWS. The SIM card is the most secure token in mass production at the moment and can easily store all the sensitive information needed (certificates, keys, personal information). After all, even if a phone or SIM card is lost or stolen, it will be quite difficult for someone to extract the necessary information and by the time that he manages to do so, the certificates/keys will, most probably, be revoked.

A third reason is the transparency of the process. As mentioned earlier, the user needs to know how to use a browser and nothing more. All the necessary message exchange takes place without user interaction except when entering the PIN number and this allows for more complex protocols and longer cryptographic keys to be used.

The most difficult part of the overall process is when it comes to define who will be responsible for storing all these certificates/keys on the SIM card. Are the mobile operator companies trusted to install this sensitive information on the SIM cards, and in case they are not, will they allow the government to use their facilities? What happens with lost/stolen phones (revoking of the certificates), or simply when a user changes phone and/or network provider?

Additionally, it must be ensured that everything runs smoothly, so that the election result is not disputed and that the legitimate user can vote only once. This is quite a challenge as although the Internet is used by more and more people for all kinds of different purposes, it is far from being characterized as a secure environment. The ability that it provides for shopping, communicating, etc. intrigues malicious people and offers them a whole new environment where they can launch their attacks against unwary users. A number of malicious programs are created every day, including Trojan horses, viruses, rootkits and other attack software that is used for data theft, communications corruption, information destruction, etc. Although the SIM card is designed as an attack/tamper-resistant platform, extending its ability to serve HTTP(s) requests will make the SIM card and mobile phone even more attractive attack targets.

19.3.3 Secure Communication Channel

There are two ways for the SCWS to communicate with entities outside of the SIM card environment. The first way is when the communication is between the SCWS and a remote server in order for the former to receive updates and the second one is the communication with the phone's browser when the user submit requests to the SCWS. Both communication mechanisms need to be protected adequately.

The communication between the SCWS and the remote server is of vital importance, because it provides the necessary updates to the SCWS from a central location. The symmetric cryptography can provide the necessary level of security through the use of a pre-shared symmetric key [8]. The key has to be strong (long) enough so that even if the communication is eavesdropped, an attacker cannot decrypt or alter it. In addition, this offers mutual authentication, because the key is only known to the two entities, thus every message encrypted by that key can only come from a trusted entity.

The security of communication between the SWCS and the browser is also very important and the necessary level of protection can be provided with in a variety of ways. As with the traditional Internet, the user can either use the HTTP or the HTTPs protocols to communicate with the SCWS. If the browser on the mobile phone requests information that needs little or no security, the communication can pass over the HTTP protocol, while communication that is sensitive is protected via HTTPs, thus offering confidentiality, integrity and authentication. Another security measure that can offer a second level of security is the use of the PIN to authenticate the user to the SCWS. The final measure that can be utilized is through the use of some form of ACL that will allow applications meeting certain trust criteria, to access and communicate with the SCWS, while blocking non trusted applications.

19.3.4 Data Confidentiality/Integrity

The SCWS handles two kinds of data: data stored on the SIM card and data in transit. The first kind of data has an adequate level of protection as modern smart cards are designed to strongly resist unauthorized access to the card data. An attacker should need costly and advanced equipment, expert knowledge and a lot of time, as modern smart cards have many countermeasures to resist known attacks [9].

Data in transit cannot benefit from the protection that the card offers and is more exposed to attacks. If data in transit does not pass over secure channels with the use of the necessary protocols, this may lead to data that is altered, destroyed or eavesdropped. Measures must be taken so that these actions are detected and if possible, prevented.

An attacker may alter data in two ways: by just destroying a message (or transforming it into a meaningless one) or by trying to produce a new version of the message with an altered, meaning. The first attack simply wants to “break” the communication, while the second one aims to exploit a weakness e.g. to execute malicious commands against the server. It is obvious that the latter is far more difficult, especially when the message is encrypted or hashed. In the SCWS context, the messages are mostly the commands exchange between the SCWS and the remote server for remote administration or between the SCWS and the user’s browser. The alteration of the exchanged messages can be avoided by adding a

MAC at the end of the message when there is a pre-shared key (in the case of the remote administration) or by using digital signatures when a pre-shared key cannot be (securely) exchanged.

Using any form of strong encryption can provide the necessary confidentiality needed for the data that is handled by the SCWS and symmetric or public-key cryptography can be used according to needs. OMA proposes the use of PSK-TLS for confidentiality/integrity between the SCWS and the remote server and public key cryptography for the communication between the SCWS and the various applications on the phone. On the second case the use of PSK-TLS is optional.

19.3.5 Authentication

For authentication purposes, OMA proposes the use of Basic Authentication and optionally the use of Digest Authentication [10]. While the former can be used when there is no or little need for authentication, in case that an application/entity needs to authenticate the SCWS and vice versa, the use of Digest Authentication is mandatory.

19.4 Management Issues

Managing a web server is a complicated task, because of all the different possibilities that exist for setting it up and tuning it. On top of that, the administrator must pay attention to its security and implement necessary countermeasures so that the server is not an easy target to possible attacks. Additionally, he must pay attention to setup correctly the scripting server-side/scripting language(s) that the web server will use in order to avoid setup mistakes that affect a large number of servers e.g. PHP's register_global problem [11]. Therefore to setup a web server correctly, an administrator must be aware of all the latest vulnerabilities, which could be quite a challenging task, especially as correcting a wrong setup option may sometimes lead to corrupted programs/websites that do not work [12].

The SCWS is not affected by these issues. Most probably it will be setup from a central location (the network operator's facilities), and must be carefully managed, because otherwise it may introduce vulnerabilities to attacks against it and against the SIM card itself. Most probably the SCWS will have a common setup for all its instances. An important problem is what to do with older SIM cards that may not be able to host a program like a web server (even a "lightweight" web server like the SCWS). This may lead to a large number of users being unable to have the SCWS installed. Finally, any patches/updates needed for the SCWS will be installed from a central location, meaning that this adds another burden to the network provider.

19.5 Personalization

At the beginning, the Internet was a static environment with content that was presented to the user "as is", without interaction. However, since the arrival of Web 2.0 this has changed radically; now it is often that the user "creates" and personalizes the content [13]. Social networking sites, blogs and other internet sites allow a user to be an author, to present his photographic skills, to communicate with people from all over the world, to create the content in general. This advance in Internet interactivity, allowed a number of companies, to approach the user offering services or products that were of interest according to Internet "habits" e.g. a person that visits sports sites would be more interested in sports clothing than a person that visits music sites.

The idea of the personalized content can be applied to the SCWS environment as well. Network companies may offer services that interest their clients based not only on their needs/interests, but e.g. if a client is in a different country, the company may provide information about that place (museums, places of interest, hospitals, other useful local information) and send this information to the SCWS. The user can then easily, using his phone browser, access the SCWS content, even if the user is offline (not connected to the internet).

One issue is to define who will be the creator of the content. Will this be the network provider or will third parties be allowed to offer content as well? Giving access to a third party may be resisted for business reasons and also the potential for undermining the security of the platform. From a practical viewpoint, multiple applications are not too much of a challenge for the modern SIM card platform, as it is designed to permit third parties to install, manage and run applications.

19.6 Web Server Administration

The administration of a web server can be quite a challenging task, due to the fact that the server must run smoothly, work 24/7/365 and serve from a few hundred to even million requests (depending on the sites that is hosts).

The SCWS will not be installed in a central location like a traditional web server, but rather it will be installed in a (large) number of phones. The web server can come pre-installed on the SIM card, and when an update/patch must take place this can happen with one of the following two ways: either centrally with mass distribution of newer versions/updates or by presenting a page to the user, for self download and install. Third party applications that may be installed and run as part of the web server can be administered in the same way.

19.7 Web Server's Processing Power and Communication Channel Speed

Depending on the need, a web server can be slow or fast. A server that has to support a hundred requests does not need the same bandwidth or the processing power as one that serves a million requests. The supporting infrastructure is of huge importance, so that users' requests are handled swiftly even in the case of some server failures. Additionally, if the performance of the communications bearer does not match the processing power of the web server itself (and vice versa), the user's perspective of the overall performance will be poor.

The SCWS will not serve requests for more than a user (if we assume that the SCWS is only accessible from inside the phone), so one can say that processing power or the communication channel is not of huge importance. However, even if the SCWS has to respond to only one request at a time, this may still be a demanding task if the processing power of the SIM card is still small, especially if the SCWS has to serve multimedia or other resource consuming content.

The same concern applies to the communication channel speed. A traditional web server may have a fast line (or maybe more than one) along with backup lines to serve its requests. The SCWS cannot only rely on the traditional ISO 7816 interface [14]. This interface that exists in most of the phones at the moment is too slow to serve incoming and outgoing requests to and from the SCWS. This need is recognized by ETSI and it is expected in the near future that the 7816 interface will be replaced with the (much) faster USB one [15]. So, the necessary speed can be provided only when the USB interface is widely available.

19.8 IP Mobility

A web server that changes IP addresses is a web server that may not be as accessible as it must be, because every time that its IP address is changed a number of DNS servers must be informed and their databases to be updated. This update may require from several hours to a couple of days. This is the reason that all web servers have static IP addresses [16], so that every time that a user enters a URL, he knows that there will be a known match between that URL and an IP address.

At the beginning the SCWS will not be as accessible as a traditional web server. This is because (as mentioned before) it will serve requests initiated from one user only, as it will be accessible only from the inside of the phone and most probably its IP address will be the 127.0.0.1 (loopback address). However, if the SCWS becomes accessible to entities from outside of the phone, this means that it cannot answer to requests destined to the loopback address only and it will need a public accessible address. Although this can be solved and each SCWS can have a static IP address, what will happen when the user is travelling? After all, IP address ranges are assigned to cities or countries and so a PC has an IP within a certain range when being in London, UK and another when being in Thessaloniki, Greece.

A phone is a mobile device which is often transferred between cities, countries or even continents and so it needs a different IP address every now and then. This means that if the SCWS becomes public accessible and serves requests to entities from outside the phone, there must be a way to permanently match the SWCS's URL to a certain IP. While this cannot happen because we are talking about a mobile device, the answer to this problem can be found within RFC 3344 and RFC 3775 (for mobile IPv4 and IPv6 respectively). Briefly, the two RFCs use a home address and a care-of-address. The packets that are destined for the home address are forwarded to its care-of-address (which is the phone's current address). This binding requires co-operation between the network providers, but if it is setup correctly it can enable the phone to move without problems and permit the SCWS to serve requests smoothly [17, 18].

19.9 Conclusion

To predict the future of the SCWS is not an easy task, but it surely may be interesting. Obstacles associated with the SIM cards' limited processing power and low bandwidth communication channel may be overcome by advances in current technology. Newer SIM cards have bigger memory capabilities; faster processing units and the USB interface will provide the necessary communication speed. However, technical issues alone are unlikely to decide the future of SCWS, as this will be primarily determined by the network providers based on profitability, potential security vulnerabilities and user acceptance.

References

1. Rees J, Honeyman P (1999) Webcard: a java card web server. Center for Information Technology Integration, University of Michigan
2. Open Mobile Alliance. http://www.openmobilealliance.org/
3. OMA, Enabler Release Definition for Smartcard-Web-Server Approved Version 1.0-21 April 2008, OMA-ERELD-Smartcard_Web_Server_V1_0-20080421-A
4. OMA, Smartcard Web Server Enabler Architecture Server Approved Version 1.0-21 April 2008, OMA-AD-Smartcard_Web_Server_V1_0-20080421-A
5. OMA, Smartcard-Web-Server Approved Version 1.0-21 April 2008 OMA-TS-Smartcard_Web_Server_V1_0-20080421-A
6. ETSI TS 102 223
7. AS Sertifitseerimiskeskus, The Estonian ID Card and Digital Signature Concept Principles and Solutions, Version 20030307
8. Menezes AJ, van Oorschot PC, Vanstone SA (1996) Handbook of applied cryptography. CRC Press, USA, pp 15–23, 352–359
9. Rankl W, Effing W (2003) Smart card handbook, 3rd edn. Wiley, New York, pp 521–563
10. Internet Engineering Task Force, HTTP Authentication: Basic and Digest Access Authentication. http://tools.ietf.org/html/rfc2617
11. PHP Manual, Using Register Globals. http://php.net/manual/en/security.globals.php

12. Esser S, $GLOBALS Overwrite and its Consequences. http://www.hardened-php.net/globals-problem (November 2005)
13. Anderson P (2007) What is Web 2.0? Ideas, Technologies and Implications for Education. Technology & Standards Watch, February 2007
14. Mayes K, Markantonakis K (2008) Smart cards, tokens, security and applications. Springer, Heidelberg, pp 62–63
15. ETSI SCP Rel.7
16. Hentzen W, DNS explained. Hentzenwerke Publishing. Inc, USA pp 3–5
17. Internet Engineering Task Force, IP Mobility Support for IPv4, http://www.ietf.org/rfc/rfc3344.txt
18. Internet Engineering Task Force, Mobility Support in IPv6, http://www.ietf.org/rfc/rfc3775.txt

Chapter 20
A Scalable Hardware Environment for Embedded Systems Education

Tiago Gonçalves, A. Espírito-Santo, B. J. F. Ribeiro and P. D. Gaspar

Abstract This chapter presents a scalable platform designed from scratch to support teaching laboratories of embedded systems. Platform's complexity can increase to offer more functionalities in conjunction with student's educational evolution. An I2C bus guarantees the continuity of functionalities among modules. The functionalities are supported by a communication protocol presented in this chapter.

20.1 Introduction

Embedded systems design plays a strategic role from an economic point of view and industry is requiring adequately trained engineers to perform this task [1, 2]. Universities from all over the world are adapting their curriculums of Electrical Engineering and Computer Science to fulfill this scenario [3–7]. An embedded system is a specialized system with the computer enclosed inside the device that it controls. Both, low and high technological products are built following this concept.

T. Gonçalves (✉) · A. Espírito-Santo · B. J. F. Ribeiro · P. D. Gaspar
Electromechanical Engineering Department, University of Beira Interior, Covilhã, Portugal
e-mail: blueflash.pt@gmail.com

A. Espírito-Santo
e-mail: aes@ubi.pt

B. J. F. Ribeiro
e-mail: bruno@ubi.pt

P. D. Gaspar
e-mail: dinis@ubi.pt

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_20,
© Springer Science+Business Media B.V. 2011

The complexity of an embedded system changes from one product to another, depending on the task that it must perform. Therefore, embedded system designers must have knowledge from different areas. The development of hardware projects requests knowledge related with digital and/or analog electronics, and, at the same time, with electromagnetic compatibility issues, that cannot be forgotten in high frequency operation or in products that must work in very restrictive environments, as the ones found in hospitals. Therefore, the designer must project the firmware, required by the hardware, allowing it to work as expected. Subjects like operating systems, real time systems, fixed and floating-point arithmetics, digital signal processing, and programming languages as assembly, C/C++ or Java are of major relevance to the development of embedded systems.

Curriculums to teach embedded systems are not well established, unlike the classic knowledge areas where is possible to find textbooks to support students study. Different sensibilities are used to structure curriculums in embedded systems. The ARTIST team has established the competencies required by an embedded software engineer. This proposal highlights practices as an essential component of education in embedded system [8].

As stated previously, the skills that an embedded systems designer must hold are highly complex and spread across different areas [9, 10]. If beyond this knowledge, the student needs also to learn how to work with a complex development kit, then he/she will probably fail. Even if the student is adequately prepared in the essential subjects, the time taken to obtain visible results is, sometimes, responsible for student demotivation and consequent failure.

20.2 Hardware Platform Design Overview

The development of an embedded system relies on a large and assorted set of technologies in constant and rapid evolution. The learning platform here described wish to improve student accessibility to this set of technologies in an educational environment [11]. This way, the usage of commercial evaluation kits are discouraged, since they are mainly developed to observe the potentialities of a specific product without educational concerns.

The MSP430 family selection is justified by the amount of configurations available, with a high count of peripherals, different memory sizes, and popularity. Another relevant attribute is the learning curve of these devices that, from authors' experience, allows to rapidly obtain results [12].

The developed platform has actually four modules (see Fig. 20.1), with increasing complexity, but others can be developed in the future. The learning platform can thus satisfy the needs from beginners to experienced students, and, at the same time, has a high potential of evolution.

The learning platform here presented allows users to experiment with different kinds of interfaces, such as OLED display, seven-segment displays and conventional LED. Students can interact with several other peripherals, as for example, among others: pressure switches, a touch-screen pad, a joystick, and an

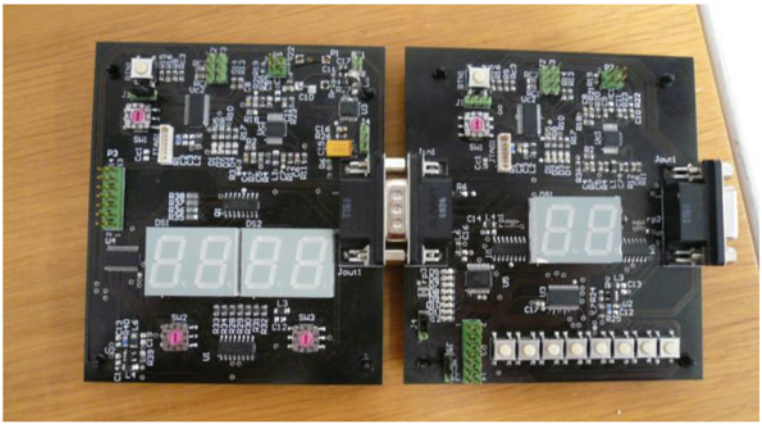


Fig. 20.1 Example of one allowed configuration—Module 0 and Module 1 connected

accelerometer. An overview structure of the developed modules can be observed in Fig. 20.2. The modules were constructed around three devices from the MSP430 family: MSP430F2112, MSP430FG4618 and MSP430F5419.

The strategy adopted to design the teaching platform took in consideration the following characteristics:

- Modules can exchange data through an I2C bus.
- Each module has two microcontrollers, one manage the communications, while the other it is available for student's work.
- A SPI bus connects both processors in the same module.
- Energy consumed in each module can be measured and displayed in real time.
- All modules have standard dimensions.

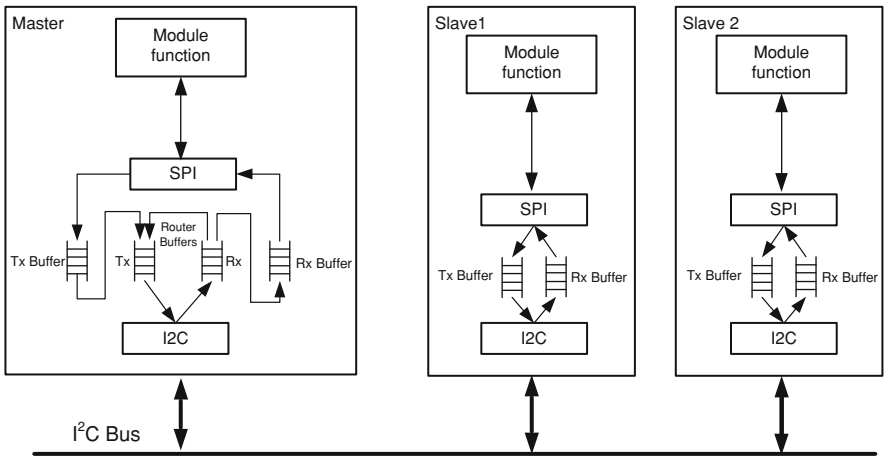


Fig. 20.2 Structure overview of the teaching environment for embedded system

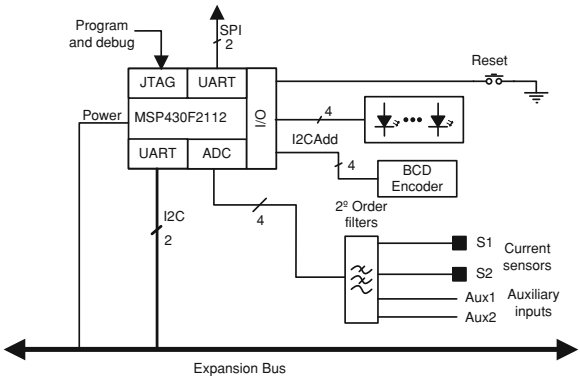


Fig. 20.3 Communication infrastructure

The communication management infrastructure shown in Fig. 20.3 implements the protocol described in Sect. 20.4. All the functionalities specified for the Module 0—Basic Interface and Power—and for the Module 1—Basic Interface—are implemented with the MSP430F2112 microcontroller. This device can operate with a maximum clock frequency of 16 MHz, it has 32 kB of flash memory, and 256B of RAM, one 10-bits ADC, two timers with respective compare/capture units, and a number of digital IO sufficient to satisfy the needs. The device also supports SPI, UART, LIN, IrDA, and I2C communication protocols.

A BCD encoder sets the address of the module in the I2C bus. Four LED are used to show the communication status. A pressure switch can reset the communication management hardware. An SPI channel connects the user's microcontroller and the microcontroller used to manage the communications, which in turn is connected to the I2C bus.

The MSP430 family has good performance in situations where the energy consumption is a major concern. The energy consumed by each module can be monitored and displayed in real time. Current is measured at two distinct points. At measuring point S1 the current of the user's microcontroller is acquired with a current shunt monitor (INA198). The total current required by the whole module is acquired at the measuring point S2 with a precision resistance.

Two other analogue inputs are also available. The student can use them to acquire the working module voltage. The energy consumed is computed with the knowledge of the current and voltage.

20.3 Platform Modules in Detail

20.3.1 Module 0: Basic Interface and Power

This module was developed to help students' first steps in embedded systems. Usually, this kind of user does not have any experience with the development of embedded systems. This module will help him to take contact with the

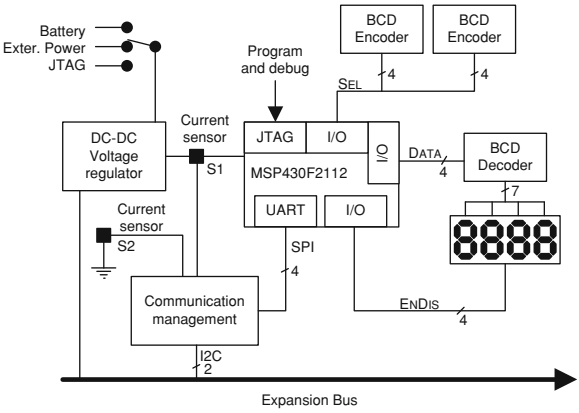


Fig. 20.4 Module 0—structure of the basic interface and power module

microcontroller architecture and, at the same time, with the software development tool. The internal structure of the module is illustrated in Fig. 20.4.

Beyond providing power to itself, the module can also power the modules connected to it through the Expansion Bus. Three different options are available to power the system: a battery, an external power source, or the JTAG programmer. The DC–DC converter allows an input range from 1.8 to 5.5 V, providing a 3.3 V regulated output. Powering from the JTAG is only available to support programming and debugging activities.

A numerical display was built with four seven-segment independent digits. Four data lines (DATA) control the writing operation. A BCD decoder allows writing the desired value in the display. The selection of which digit will be written, at a specific moment, is performed by four control lines (ENDIS). Because the BDC decoder does not latch the output, the microcontroller must continually refresh the value to exhibit in the display with a minimal frequency of 15 Hz.

Despite the simplicity of this module, its predefined task is the visualization of the current and the energy consumed by each one of the modules connected by the Expansion Bus. Two BCD encoders, with ten positions each, are connected to the microcontroller through eight selection lines (SEL), and are used to select from which module the information will come. To execute this feature, the user's microcontroller must be loaded with a specific firmware.

20.3.2 Module 1: Basic Interface

This module is directed to the student that already has some basic knowledge in the embedded systems field. As the previous module, this is also based in the MSP430F2112 microcontroller. Connected to this device, as can be seen in Fig. 20.5, can be found eight switches, eight LEDs, and a seven-segment display with two digits.

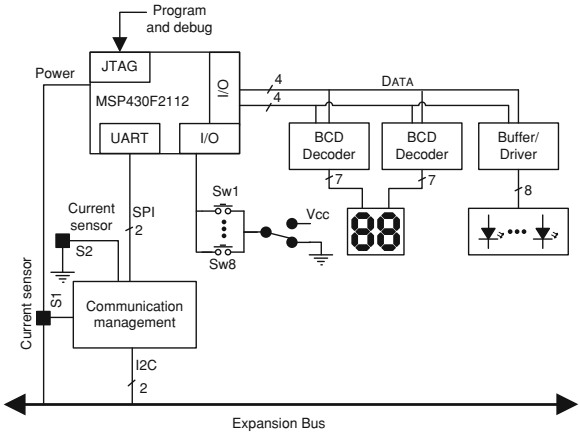


Fig. 20.5 Module 1—structure of the basic interface module

This module intends to develop student's competences related with synchronous and asynchronous interruptions. Simultaneously, the student can also explore programming techniques, as for example, the ones based in interrupts or port polling to check the status of the digital inputs or impose the status of the digital outputs. With this module, the student can have the first contact with the connection of the microprocessor to other devices, compelling him to respect accessing times. Two BCD decoders, with the ability to latch the outputs, are used to write in the display. While the DATA lines are used to write the value in the display, the LE lines are used to select which digit will be written. The DATA data lines are also used, with a buffer/driver, to turn on or off each one of the eight LED. The user can configure the inactive state of the switches.

20.3.3 Module 2: Analog and Digital Interface

Students with more advanced knowledge in the embedded systems field can use this module to improve their capabilities to develop applications where human–machine interface, analogue signal conversion, and digital processing are key aspects. The design of this module was performed around the MSP430FG4618 microcontroller. This device can operate at maximum clock frequency of 8 MHz, it has 116 kB of flash memory and 8 kB of RAM. The high count of digital IO is shared with the on-chip LCD controller. Other peripherals that are normally used for analog/digital processing are also present: 12-bits ADC, two DAC, three operational amplifiers, DMA support, two timers with compare/capture units, a high number of digital IO, and hardware multiplier. This device can also support SPI, UART, LIN, IrDA, and I2C communication protocols. On-chip LCD controller allow the connection of LCD with 160 segments. On-board can be found a navigation joystick with four positions and a switch, a rotational encoder with 24 pulses/turn and a switch, a speaker output, a microphone input, generic IO, and a

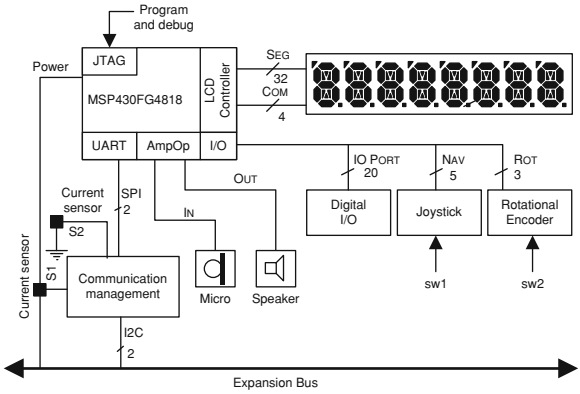


Fig. 20.6 Module 2—structure of the analog and digital interface module

alphanumeric LCD. The internal structure of the module 2—Analog and digital interface—is illustrated in Fig. 20.6.

Students can explore the operation of LCD devices, taking advantage from the on-chip LCD controller. The module can also be used to improve the knowledge related with the development of human–machine interfaces.

An example of a laboratory experience that students can perform with this module is the acquisition of an analogue signal, condition it with the on-chip op-amps, and digitally process the conversion result with a software application. The result can be converted again to the analog world using the on-chip DAC.

Taking advantage from the on-chip op-amps it is possible to verify the work of different topologies, as for example: buffer, comparator, inverter non-inverter, and differential amplifier with programmable gain.

The digitalized signal can be processed using the multiply and accumulate hardware peripheral. Students can conclude about the relevance of this peripheral in the development of fast real time applications.

20.3.4 Module 3: Communication Interface

This is the most advanced module. With this module, the student has access to a set of sophisticated devices that are normally incorporated in embedded systems. The student can explore how to work with: an OLED display with 160 × 128 pixels and 2,62,000 colors, a three axis accelerometer, a SD card, a touch-screen, an USB port, two PS/2 interfaces. The module also owns connectors to support the radio frequency modules Chipcon-RF and RF-EZ430.

The Module 3—Communications interface module—illustrated in Fig. 20.7, was built around a microcontroller with high processing power. The MSP430F5419 can operate with a maximum clock frequency of 18 MHz, and it has 128 kB of flash memory and 16 kB of RAM. This microcontroller also has a hardware multiplier,

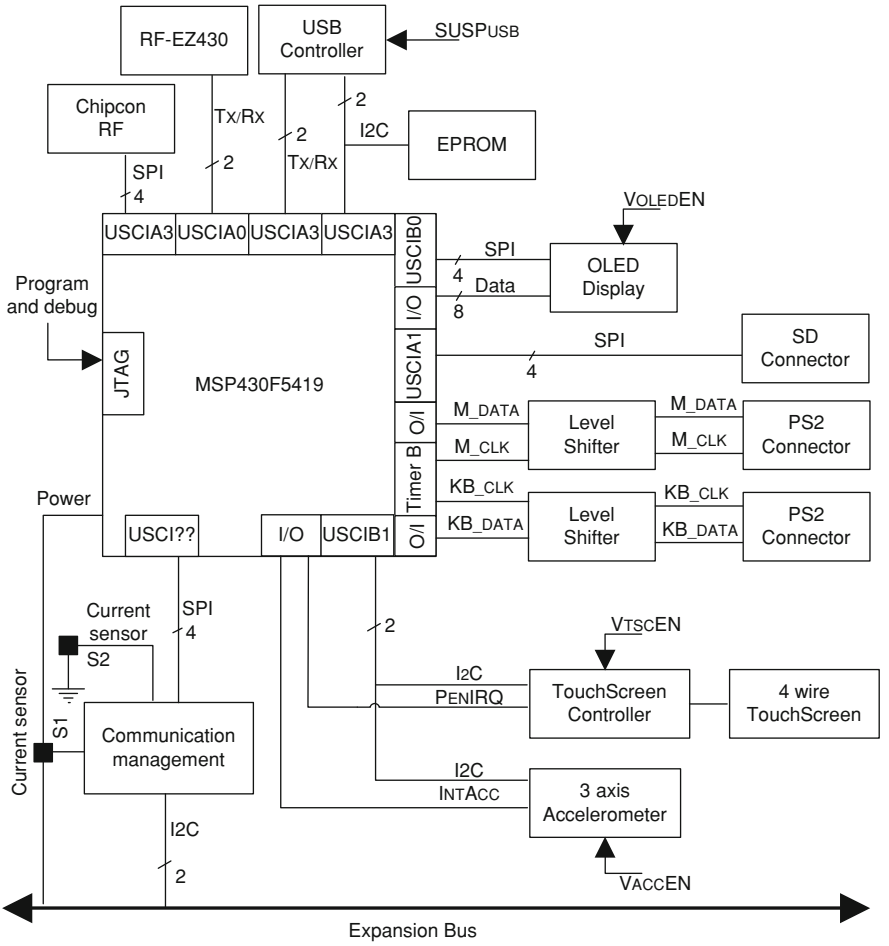


Fig. 20.7 Module 3—structure of the communication interface module

a real time clock, direct memory access, and a 12-bits ADC. The communication protocols SPI, UART, LIN, IrDA, and I2C can be implemented in four independent peripherals. The device has three independent timers with compare/capture units.

The accelerometer MMA7455L, with adjustable sensibility, uses an I2C interface to connect with the microcontroller in the port USCIB1. It has two outputs that can signal different conditions, like data available, free fall, or motion detection. This device can be enabled by the microcontroller through the VACCEN line.

The resistive touch-screen with four wires, and a 6×8 cm area, needs a permanent management of their outputs. To free the microcontroller from this task, a touch-screen controller is used to connect it with the microcontroller through I2C bus (USCIB1). The line PENIRQ notifies the microcontroller that the

touch-screen is requesting attention. To save power, the microcontroller can enable or disable the touch-screen controller through the VTSCEN line. The two PS/2 ports are connected to the microcontroller by a bidirectional level shifter to adapt the working voltage levels from 5 to 3.3 V. The clock signal is provided by the Timer B. The communication port USB uses a dedicated controller that makes the interface with the UART (USCIA2). The USB controller firmware is saved in an EPROM that can communicate with the USB controller or with the microcontroller by an I2C bus (USCIB3). The USB can be suspended or reseted by the microcontroller though the line SUSPUSB. The SD socket is only a physical support to allow the connection of the card to the microcontroller through a SPI (USCIA1). Two additional lines enable detection and inhibition of writing operations. The OLED display support two different interface methods: 8-bits parallel interface, or SPI (USBI0). The parallel interface requests a specific software driver. A simpler interface can be implemented with the SPI bus. The radio frequency interfaces allow the connection of a RF module from the Chipcon, connected to the microcontroller by an SPI interface (USCIA3). The RF-EZ430 radio frequency module can be connected through an UART (USCIA0).

20.4 Communications Protocol

Data exchange between modules is based upon two communication methods: (1) the communication method between the module application function and the network function; and (2) the communication method supported by the bus, which has the main task of interconnecting all modules, working as router among them. This characteristic gives to the system a high versatility, allowing increasing the set of applications supported.

The adopted solution was a serial bus, being the key criteria of this choice: the physical dimensions, the maximum length of the bus, the transmission data rate and, of course, the availability of the serial communications interface. The I2C communication protocol is oriented to master/slave connection, i.e., the exchange of data will always occur through the master. This leads to the definition of two distinct functional units. Although, this technology supports multi-master operation, it was decided to use a single master, giving to the network a hierarchical structure with two levels.

All the possibilities of information transaction at the bus physical level between the master and the slave are represented in Fig. 20.8. The master begins the communication sending a start signal (S), followed by the slave address (Adress X). The slave returns the result to the master after the reception and execution of the command. The procedure to send and receive data from and to the slave is also represented. The master starts the communication sending the start signal (S), followed by the slave address. The kind of operation to perform (read or write) is sent next. The ACK signal sent by the I2C controllers is not represented in the figure.

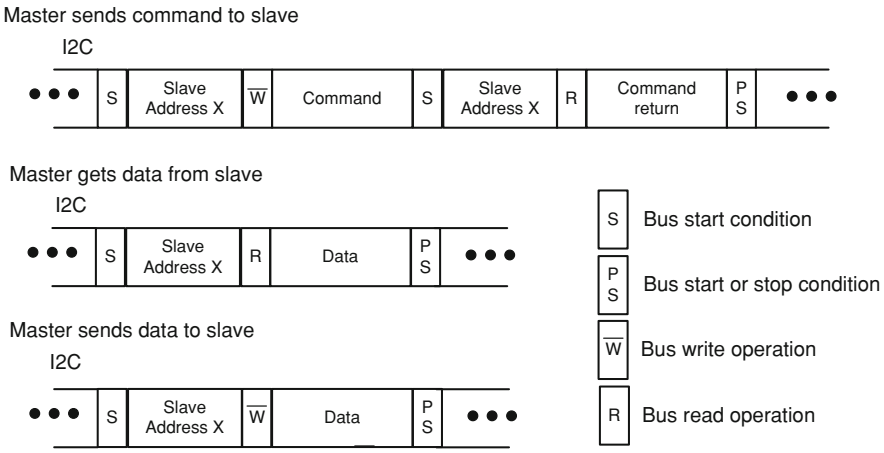


Fig. 20.8 Typical master/slave data exchange at the I2C bus level

The module 0—Basic interface and power—performs the master role. This choice is justified by the fact that this module has an obligatory presence in the network. If slave unit B wants to send a message to the slave unit A, first the message must be sent to the master, which will resend it to the destination unit. To satisfy the specifications, a communication frame was defined as can be observed in Fig. 20.9. In order to implement the communication service layer, a media access protocol was outlined with the mechanisms required to exchange information between modules. Therefore, the service frame and the communication protocols between the master unit and the slaves are defined as follows.

The service frame has three different fields. The header field, with two bytes length, has the information about the message type. The routing information includes sender and receiver addresses, and information about the length of the payload in bytes. The payload field is used to carry the information from one module to another. Finally, the checksum field, with one-byte length, controls communications integrity. The service layer uses a command set with three main goals: network management and setup; information transaction tasks synchronization; and measurement.

The service layer protocol has three different types of frames. The data frame is used to transport the information between the applications running in the modules. The command frame establishes the service layer protocol. The message type is specified in the sub-field frame type of the header field as reported in Table 20.1. The command set is listed in Table 20.2.

At power on, the master will search for slaves available to join the network. The master achieves this task sending the command “Get_ID request” for all available slave address, which will be acknowledged by the slaves present responding the command “GET_ID response”.

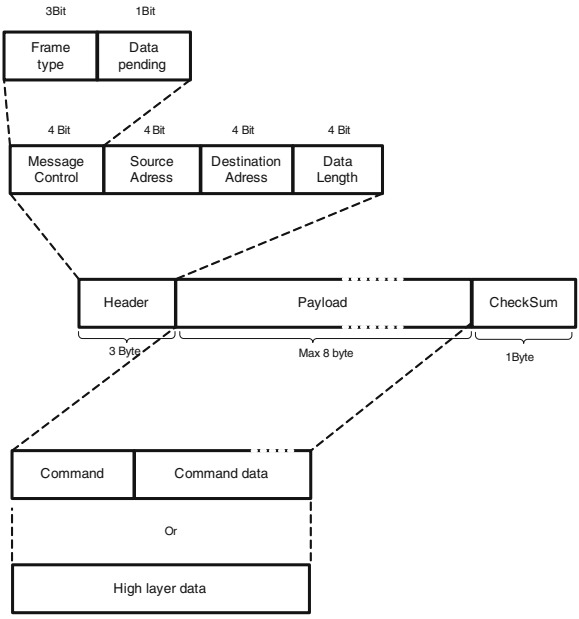


Fig. 20.9. General frame format

Table 20.1 Values of the frame type subfield

Frame type value b2 b1 b0	Description
000	Reserved
001	Data
010	Acknowledgment
011	MAC command
100–111	Reserved

Table 20.2 Command frame

Command frame identifier	Command name	Direction
0x00	Heart _Beat Request	M -> S
0x01	Heart_Beat Response	S -> M
0x02	Get_ID Request	M -> S
0x03	Get_ID response	S -> M
0x04	Software Reset Request	M -> S
0x05	Get_Status Request	M -> S
0x06	Req_Data Request	M -> S
0x07	Req_Data Response	S -> M

After the successful setup of the network, the master periodically executes the data polling operation. The task checks all slaves for pending messages. This operation allows data exchange among all the units connected by the I2C bus. The memory available in the master to support the communication task is limited,

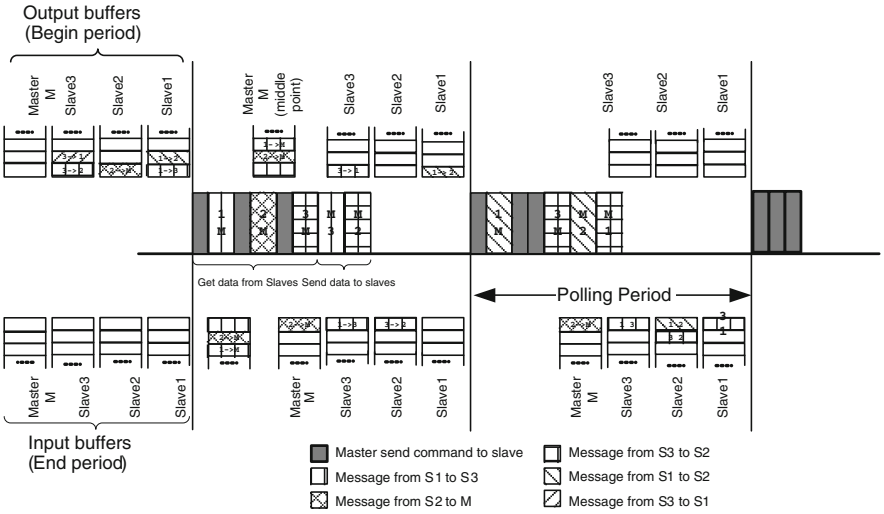


Fig. 20.10 Polling data operation

furthermore, the data polling operation must ensure low access times to the bus for all units connected to it. The pooling operation is illustrated in Fig. 20.10, where the data flow through the bus is represented.

The data polling operation is carried out in three different phases. (i) Slaves are sequentially scanned by the master searching for messages ready to be transferred. (ii) During the router phase the master inspects the field address from each message in the input buffer, if the message has a slave as the destination then it will be transferred to the output buffer, if the message has the master as the destination, then it will be sent to the user's microcontroller. (iii) The data polling operation is finished by the master sending all messages in the output buffer to the respective slaves.

20.5 Conclusion

This chapter presents the development of a platform to support the teaching of embedded systems. The modules here presented allow the implementation of different experimental laboratories, with increasing level of difficulty. At the same time, the student has access to key technologies related with the development of embedded systems.

The presence of an expansion bus gives a high versatility to the learning platform, because, it allows the development of new modules. This versatility is further enhanced by the existence of a communication bus that turns possible data exchange between modules.

References

1. Choi SH, Poon CH (2008) An RFID-based anti-counterfeiting system. IAENG Int J Comput Sci 35:1
2. Lin G-L, Cheng C-C (2008) An artificial compound eye tracking pan-tilt motion. IAENG Int J Comput Sci 35:2
3. Rover DT et al (2008) Reflections on teaching and learning in an advanced undergraduate course in embedded systems. IEEE Trans Educ 51(3):400
4. Ricks KG, Jackson DJ, Stapleton WA (2008) An embedded systems curriculum based on the IEEE/ACM model curriculum. IEEE Trans Educ 51(2):262–270
5. Nooshabadi S, Garside J (2006) Modernization of teaching in embedded systems design—an international collaborative project. IEEE Trans Educ 49(2):254–262
6. Ferens K, Friesen M, Ingram S (2007) Impact assessment of a microprocessor animation on student learning and motivation in computer engineering. IEEE Trans Educ 50(2):118–128
7. Hercog D et al (2007) A DSP-based remote control laboratory. IEEE Trans Ind Electron 54(6):3057–3068
8. Caspi P et al (2005) Guidelines for a graduate curriculum on embedded software and systems. ACM Trans Embed Comput Syst 4(3):587–611
9. Chen C-Y et al (2009) EcoSpire: an application development kit for an ultra-compact wireless sensing system. IEEE Embed Syst Lett 1(3):65–68
10. Dinis P, Espírito-Santo A, Ribeiro B, Santo H (2009) MSP430 teaching ROM. Texas Instruments, Dallas
11. Gonçalves T, Espírito-Santo A, Ribeiro BJF, Gaspar PD (2010) Design of a learning environment for embedded system. In: Proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July, 2010, London, UK, pp 172–177
12. MSP430™16-bit Ultra-Low Power MCUs, Texas Instruments. http://www.ti.com

Chapter 21
Yield Enhancement with a Novel Method in Design of Application-Specific Networks on Chips

Atena Roshan Fekr, Majid Janidarmian, Vahhab Samadi Bokharaei and Ahmad Khademzadeh

Abstract Network on Chip (NoC) has been proposed as a new paradigm for designing System on Chip (SoC) which supports high degree of scalability and reusability. One of the most important issues in an NoC design is how to map an application on NoC-based architecture in order to satisfy the performance and cost requirements. In this paper a novel procedure is introduced to find an optimal application-specific NoC using Particle Swarm Optimization (PSO) and a linear function which considers communication cost, robustness index and contention factor. Communication cost is a common metric in evaluation of different mapping algorithms which have direct impact on power consumption and performance of the mapped NoC. Robustness index is used as a criterion for estimating fault-tolerant properties of NoCs and contention factor highly affects the latency, throughput and communication energy consumption. The experimental results on two real core graphs VOPD and MPEG-4 reveal the power of proposed procedure to explore design space and how effective designer can customize and prioritize the impact of metrics.

A. R. Fekr (✉) · M. Janidarmian
CE Department, Science and Research Branch, Islamic Azad University, Tehran, Iran
e-mail: a.roshan@srbiau.ac.ir

M. Janidarmian
e-mail: jani@srbiau.ac.ir

V. S. Bokharaei
ECE Department, Shahid Beheshti University, Tehran, Iran
e-mail: v.samadi@sbu.ac.ir

A. Khademzadeh
Iran Telecommunication Research Center, Tehran, Iran
e-mail: zadeh@itrc.ac.ir

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_21,
© Springer Science+Business Media B.V. 2011

21.1 Introduction

Due to ever-increasing complexity of system on chip (SoC) design, and non-efficiency of electric bus to exchange data between IP cores in giga scale, the Network on Chip (NoC) is presented with more flexible, scalable and reliable infrastructure. Different mapping algorithms for NoCs are presented to decide which core should be linked to which router. Mapping an application to on-chip network is the first and the most important step in the design flow as it will dominate the overall performance and cost [1]. The main purpose of this study is to present a new method to generate a wide range of mappings with all reasonable values of communication cost. The most appropriate mapping is selected by total cost function using a linear function. The function can be customized by a designer, considering the impact of three key parameters, i.e., communication cost, robustness index and contention factor. The proposed procedure is shown in and explained in the next sections.

Albeit the proposed approach is topology-independent, it is illustrated and evaluated for 2D mesh topology as it is widely used for most mapping algorithms.

21.2 Particle Swarm Optimization as a Mapping Generator

Many mapping algorithms have been recently proposed to improve several parameters used in the NoC design. One of the most important parameters is the communication cost. There are several available mapping algorithms which are considered to minimize the communication cost. Using small hop counts between related cores will significantly drop the communication cost. Moreover, small hop counts will reduce the energy consumption and other performance metrics like latency [2]. It can be explained that reduction of hop counts can decrease the fault tolerant properties of NoC. Therefore, the optimal solution is to minimize the communication cost while maximizing the fault tolerant properties of NoC. In this paper, particle swarm optimization (PSO) algorithm is used to achieve the optimal solution.

As a novel population-based swarm intelligent technique, PSO simulates the animal social behaviors such as birds flocking, fish schooling, etc. Due to the simple concept and ease implementation, it has gained much attention and many improvements have been proposed [3]. In a PSO system, multiple candidate solutions coexist and collaborate simultaneously. Each solution, called a “particle”, flies in the problem space according to its own “experience” as well as the experience of neighboring particles. Different from other evolutionary computation algorithms, in PSO, each particle utilizes two information indexes: velocity and position, to search the problem space (Fig. 21.1).

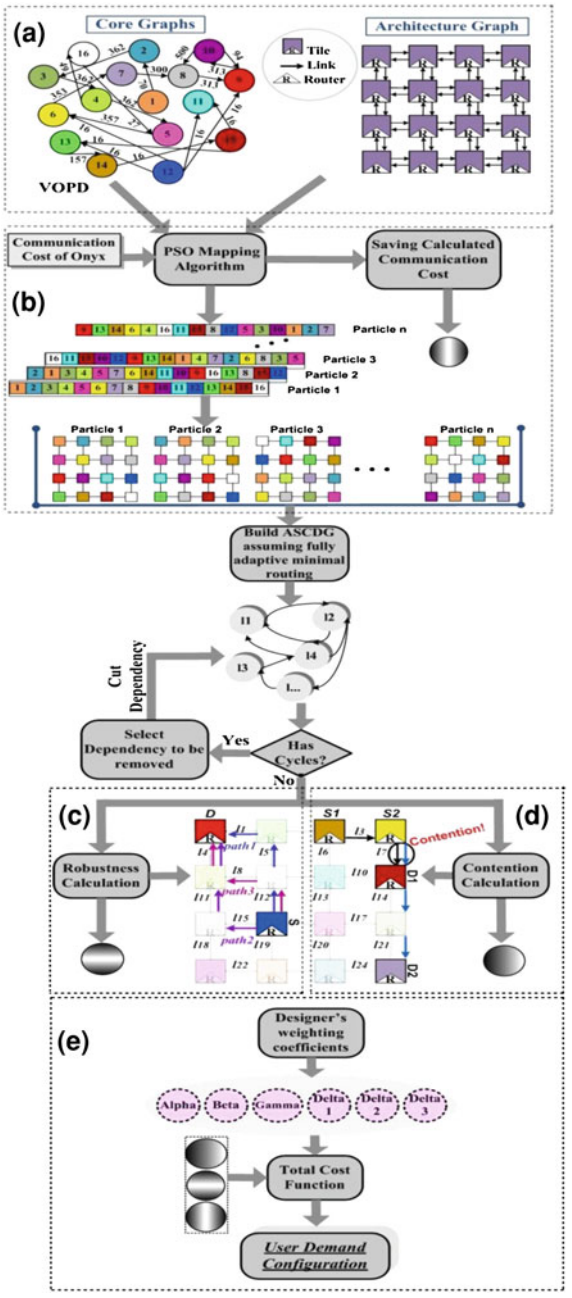


Fig. 21.1 The proposed procedure to achieve the optimal application-specific Network-on-Chip

The velocity information predicts the next moving direction, as well as the position vector is used to detect the optimum area. In standard particle swarm optimization, the velocity vector is updated as follows:

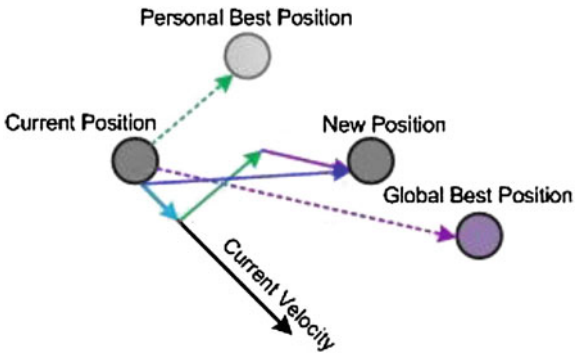


Fig. 21.2 Particle swarm optimization algorithm

$$v_{jk}(t+1) = w_t v_{jk}(t) + c_1 r_1 \left(p_{jk}(t) - X_{jk}(t)\right) + c_2 r_2 \left(p_{gk}(t) - X_{jk}(t)\right), \qquad (21.1)$$

$$w_{t+1} = w_t * w_{\text{damp}}$$

where $v_{jk}(t)$ and $x_{jk}(t)$ represent the kth coordinates of velocity and position vectors of particle j at time t, respectively. $p_{jk}(t)$ means the kth dimensional value of the best position vector which particle j had been found, as well as $p_{gk}(t)$ denotes the corresponding coordinate of the best position found by the whole swarm. Inertia weight, w_t, cognitive coefficient, c_1, and social coefficient, c_2, are three parameters controlling the size of velocity vector. r_1 and r_2 are two random numbers generated with normal distributions within interval [0,1]. With the corresponding velocity information, each particle flies according to the following rule (Eq. 21.2) [3]. This concept is shown in Fig. 21.2:

$$x_{jk}(t+1) = x_{jk}(t) + v_{jk}(t+1) \qquad (21.2)$$

It is worth mentioning that onyx is one of the best mapping algorithms in terms of communication cost. Using Onyx result and considering the evolutionary nature of PSO, different mappings are created with a variety of communication costs. To do this, onyx result is injected into population initialization step as a particle as shown in Fig. 21.1b.

In order to avoid rapid convergence, velocity threshold is not defined and c_1, c_2, w_0 and w_{damp} are set to 3.49, 7.49, 1 and 0.99 respectively in the proposed PSO algorithm. These values were obtained by examining several simulations because they drastically affect on the diversity of results.

21.3 Experimental Results of Mapping Generator

The real core graphs, VOPD and MPEG-4 [2], are used in the proposed PSO algorithm. The proposed PSO algorithm was run with 1000 initial population using 200 iterations. Figure 21.3a indicates the minimum, mean and maximum fitness function values in each iteration. As shown in Fig. 21.3b, it is clear that our PSO

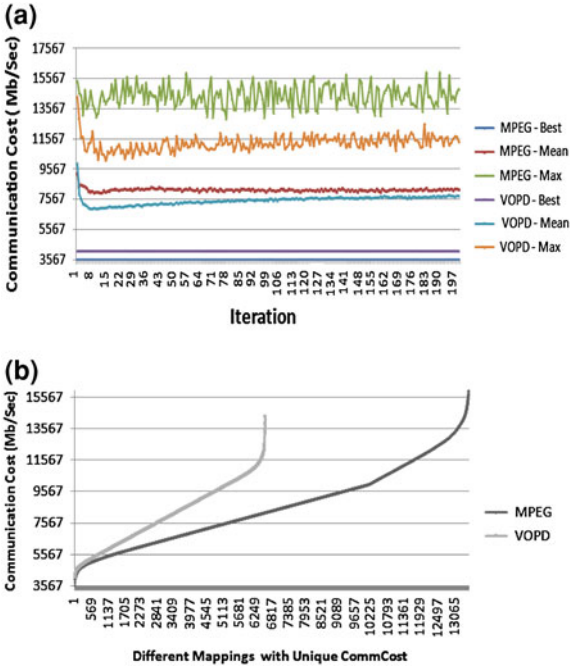


Fig. 21.3 **a** Minimum, mean and maximum fitness function values for VOPD and MPEG-4 core graphs, **b** ability of the proposed mapping generator in producing mappings with all reasonable communication cost values

algorithm could generate different mappings of VOPD and MPEG-4 core graphs with all reasonable communication cost values because of mentioned convergence control. There are 119,912 and 156,055 different unique mappings for VOPD and MPEG-4 core graphs respectively. It is worth noting that this method, which is a novel approach, enables the designer to consider other important key parameters as well.

21.4 Robustness Index

Robustness index is considered as a criterion for estimating fault tolerant properties of NoCs [4]. The greater the robustness index, the more fault tolerant NoC design. The robustness index *RI*, is based on the extension of the concept of path diversity [5]. For a given communication,$c^k \in C$, an NoC architecture graph, $A(T, L)$, a mapping function, M, and a routing function, R, [4] defined the robustness index for communication $c^k, RI(c^k)$, as the average number of routing paths available for communication, c^k, if a link belonging to the set of links used by communication c^k is faulty. Formally,

$$RI(c^k) = \frac{1}{|L(c^k)|} \sum l_{i,j \in L} \left| \rho(c^k) \backslash \rho(c^k, l_{i,j}) \right| \tag{21.3}$$

where, $\rho(c^k)$is the set of paths provided by R for communication, c^k, $\rho(c^k, l_{i,j})$ is the set of paths provided by R for communication, c^k, that uses link $l_{i,j}$, and $L(c^k)$ is the set of links belonging to paths in $\rho(c^k)$.

Suppose that there are two routing functions, A and B, which routing function A selects *path1* and *path2* and routing function B selects *path2* and *path3* to route packets between source and destination as shown in Fig. 21.1c. The routing function A selects two disjoint paths such that the presence of a faulty link in one path dose not compromise communication from source to destination since another path is fault-free. However, when the routing function B is used as shown in Fig. 21.1c, the communication will not occur. As the alternative paths share the link, l_4 any fault in the link, l_4 makes the communication from "source" to "destination" impossible. Consequently, the NoC which uses routing function A, NOC_1, is more robust than the NoC which uses routing function B, let call it NOC_2. Such situation is reflected by the robustness index. The robustness index for the above two cases are:

$$RI^{(\mathrm{NOC}_1)}(\text{Source} \rightarrow \text{destination}) = \frac{1+1+1+1+1+1}{6} = 1,$$

$$RI^{(\mathrm{NOC}_2)}(\text{Source} \rightarrow \text{destination}) = \frac{0+1+1+1+1}{5} = 0.8.$$

The NOC_1using *path1* and *path2* is more robust than the NOC_2 using *path2* and *path3* for communication from "*source*"to "*destination*"as $RI^{(\mathrm{NOC}_1)} > RI^{(\mathrm{NOC}_2)}$.

The global robustness index, which characterizes the network, is calculated using the weighted sum of the robustness index of each communication. For a communication, c^k, the weight of $RI(c^k)$ is the degree of adaptivity [6] of c^k. The degree of adaptivity of a communication, c^k, is the ratio of the number of allowed minimal paths to the total number of possible minimal paths between the source node and the destination node associated to c^k. The global robustness index is defined as Eq. 21.4.

$$RI^{(\mathrm{NOC})} = \sum_{c^k \in C} \alpha(c^k) RI^{(\mathrm{NOC})}(c^k) \tag{21.4}$$

where $\alpha(c^k)$ indicates the degree of adaptivity of communication c^k.

In this paper, one of the best algorithms which is customized for routing in application-specific NoCs, is used. The algorithm was presented in [7] which uses a highly adaptive deadlock-free routing algorithm. This routing algorithm has used Application-Specific Channel Dependency Graphs (ASCDG) concept to be freedom of dead-lock [8]. Removing cycles in ASCDG has great impact on parameters such as robustness index and is done by different methods. Therefore, in this paper, this step is skipped and left for the designer to use his preferable method.

21.5 Contention Factor

In [9] a new contribution consist of an integer linear programming formulation of the contention-aware application mapping problem which aims at minimizing the inter-tile network contention was presented. This paper focuses on the network contention problem; this highly affects the latency, throughput and communication energy consumption.

The source-based contention occurs when two traffic flows originating from the same source contend for the same links. The destination based contention occurs when two traffic flows which have the same destination contend for the same links. Finally the path-based contention occurs when two data flows which neither come from the same source, nor go towards the same destination contend for the same links somewhere in the network.

The impact of these three types of contention was evaluated and observed that the path-based contention has the most significant impact on the packet latency. Figure 21.1d shows the path-based contention. So, in this paper we consider this type of contention as a factor of mappings. More formally:

$$\text{Contention Factor} = \sum_{\forall e_{i,j} \in E} \left| L(r_{\text{map}(v_i),\text{map}(v_j)}) \cap L(r_{\text{map}(v_k),\text{map}(v_l)}) \right| \quad \text{for } i \neq k \text{ and } j \neq l \tag{21.5}$$

By having communication cost, robustness index and contention factor for each unique mapping, the best application-specific Network on Chip configuration should be chosen regarding to designer's decisions.

21.6 Optimal Solution Using a Linear Function

As previously mentioned, lower communication cost leads to an NoC with better metrics such as energy consumption and latency. Other introduced metrics were robustness index which is used as a measurable criterion for fault tolerant properties and contention factor which has the significant impact on the packet latency. A total cost function is to be introduced in order to minimize the sum of weighted these metrics (Fig. 21.1e). The total cost function is introduced as follows:

$$\text{Total Cost Function} = \text{Min}\left(\frac{\delta_1}{\alpha} \times \text{commcost}_i + \frac{\delta_2}{\beta} \times (-RI_i^{(\text{NOC})}) + \frac{\delta_3}{\gamma} \times CF_I\right) \quad \forall\ \text{mapping}_i \in \text{generated mappings} \quad \text{and } \delta_i + \delta_2 + \delta_3 = 1 \tag{21.6}$$

where, commcost$_i$ is the communication cost, $RI^{(NOC)}$ is the robustness index and CF_I is the contention factor of NoC after applying mapping$_i$.

The constants α, β and γ are used to normalize the commcost, $RI^{(NOC)}$ and CF. In this paper, α, β and γ are set to the maximum obtained values for communication cost, robustness index and contention factor. δ_1, δ_2 and δ_3 are the weighting coefficients meant to balance the metrics.

Although multi-objective evolutionary algorithms can be used to solve this problem, the proposed procedure is considered advantages in this study due to following reasons: First, a designer can change the weighting coefficients, without rerunning the algorithm. Second, due to the convergence control, the results are more diversified when compared to the multi-objective evolutionary algorithms and can be intensified by increasing the population size and/or iterations. And finally, if designer focuses on communication cost, the optimal communication cost does not usually occur in evolutionary algorithms.

21.7 Final Experimental Results

In order to better investigate the capabilities of proposed procedure shown in Fig. 21.1, we have done some experiments on real core graphs VOPD and MPEG-4. As mentioned before, one of the advantages of proposed mapping generator is its diversity of produced solutions. Based on the experimental results, mentioned mapping generator produces 201,000 mappings for VOPD and MPEG-4, according to boundaries which limit population size and maximum iteration of PSO algorithm. Dismissing the duplicate mappings led to 119,912 and 156,055 unique mappings for VOPD and MPEG-4 which extracted among whole results. Results of running this procedure for VOPD and MPEG-4 core graphs and evaluating the values in the 3D design space are shown in Figs. 21.4, 21.5, 21.6, 21.7, 21.8, 21.9, 21.10, and 21.11. Values of δ_1, δ_2 and δ_3 which used in these experiments respectively are 0.5, 0.3 and 0.2 for VOPD core graph and 0.1, 0.2 and 0.7 for MPEG-4 core graph.

As it can be seen in these figures, there are many different mappings which have the equal communication cost value that is one of the good points about proposed mapping generator. In average, there are almost 18 and 12 different mappings for each special value of communication cost while VOPD and MPEG-4 are considered as experimental core graphs. The optimal application-specific NoC configuration can be selected by setting proper values in total cost function based on designer demands. In our design, VOPD mapping with communication cost, 4347, robustness index, 54.28, and contention factor, 284, is the optimal solution. Mapping with communication cost, 6670.5, robustness index, 35.94, and contention factor, 6, is also the optimal solution for MPEG-4 mapping.

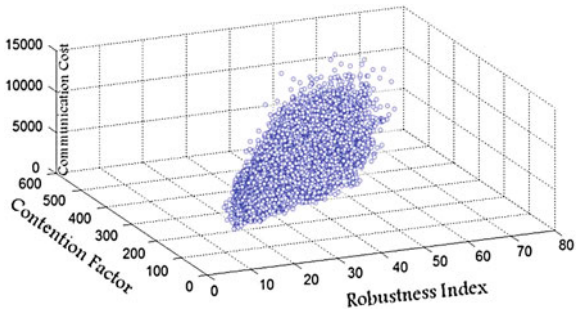


Fig. 21.4 Robustness index, contention factor and communication cost of VOPD mappings in 3D design space

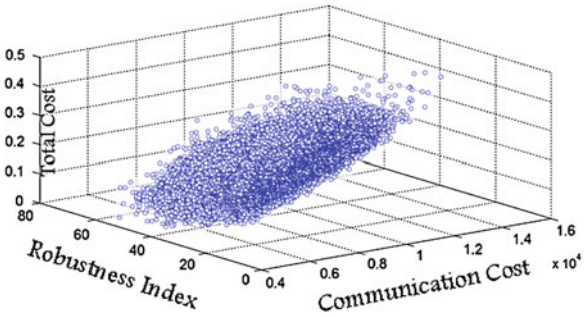


Fig. 21.5 Communication cost, robustness index and total cost of VOPD mappings in 3D design space

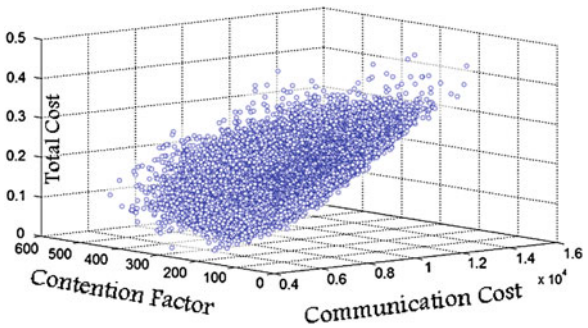


Fig. 21.6 Communication cost, contention factor and total cost of VOPD mappings in 3D design space

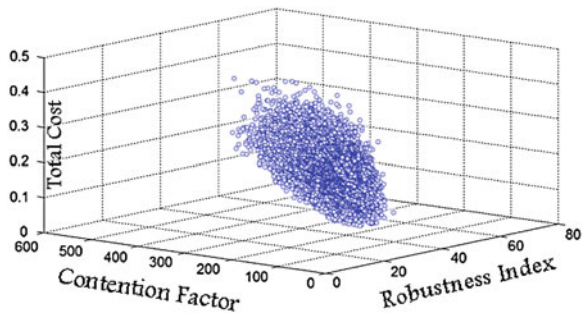


Fig. 21.7 Robustness index, contention factor and total cost of VOPD mappings in 3D design space

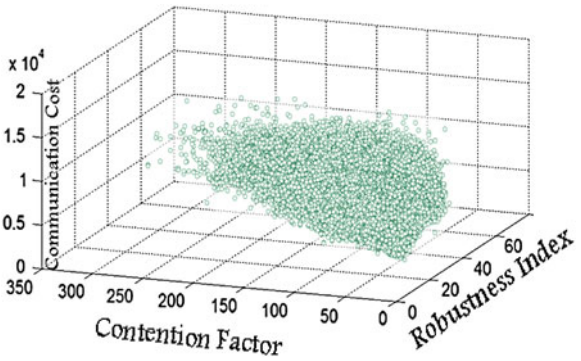


Fig. 21.8 Robustness index, contention factor and communication cost of MPEG-4 mappings in 3D design space

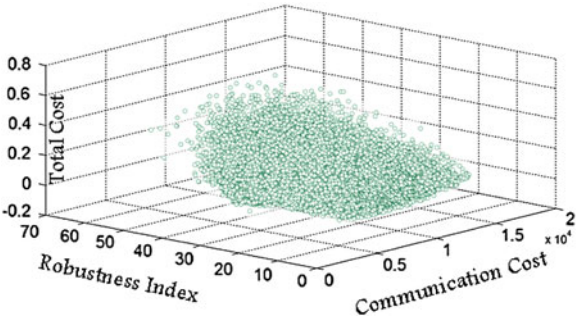


Fig. 21.9 Communication cost, robustness index and total cost of MPEG-4 mappings in 3D design space

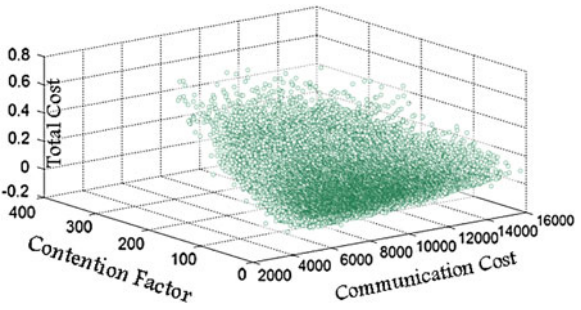


Fig. 21.10 Communication cost, contention factor and total cost of MPEG-4 mappings in 3D design space

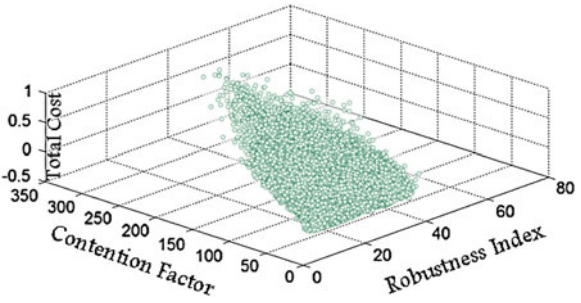


Fig. 21.11 Robustness index, contention factor and total cost of MPEG-4 mappings in 3D design space

21.8 Conclusion

As mapping is the most important step in Network-on-Chip design, in this paper a new mapping generator using Particle Swarm Optimization algorithm was presented. The best mapping in terms of communication cost was derived from Onyx mapping algorithm and injected into population initialization step as a particle. Because of using Onyx mapping results as particles, results convergence was controlled by finding appropriate values in velocity vector. This PSO algorithm is able to generate different mappings with all reasonable communication cost values. Using three metrics which are communication cost, robustness index and contention factor for each unique mapping, the best application-specific Network-on-Chip configuration can be selected regarding to designer's demands that are applied onto the total cost function.

Acknowledgments This chapter is an extended version of the paper [10] published at the proceedings of The World Congress on Engineering 2010, WCE 2010, London, UK.

References

1. Shen W, Chao C, Lien Y, Wu A (2007), A new binomial mapping and optimization algorithm for reduced-complexity mesh-based on-chip network. Networks-on-chip, NOCS, 7–9 May 2007, pp 317–322
2. Janidarmian M, Khademzadeh A, Tavanpour M (2009) Onyx: a new heuristic bandwidth-constrained mapping of cores onto tile based Network on Chip. IEICE Electron Express 6(1):1–72
3. Zhihua CUI, Xingjuan CAI, Jianchao ZENG (2009) Choatic performance-dependant particle swarm optimization. Int J Innov Comput Inf Control 5(4):951–960
4. Tornero R, Sterrantino V, Palesi M, Orduna JM (2009) A multi-objective strategy for concurrent mapping and routing in networks on chip. In: Proceedings of the 2009 IEEE international symposium on parallel & distributed processing, pp 1–8
5. Dally WJ, Towles B (2004) Principle and practice of interconnection network. Morgan Kaufmann, San Francisco
6. Glass CJ, Ni LM (1994) The turn model for adaptive routing. J Assoc Comput Mach 41(5):874–902
7. Palesi M, Longo G, Signorino S, Holsmark R, Kumar S, Catania V (2008) Design of bandwidth aware and congestion avoiding efficient routing algorithms for networks-on-chip platforms. In: Second ACM/IEEE international symposium on networks-on-chip, NoCS 2008, pp 97–106
8. Palesi M, Holsmark R, Kumar S (2006) A methodology for design of application specific deadlock-free routing algorithms for NoC systems, hardware/software codesign and system synthesis. CODES + ISSS '06. In: Proceedings of the 4th international conference, pp 142–147
9. Chou C, Marculescu R (2009) Contention-aware application mapping for Network-on-Chip communication architectures computer design, 2008. IEEE international conference on ICCD 2008, vol 19, pp 164–169
10. Roshan Fekr A, Khademzadeh A, Janidarmian M, Samadi Bokharaei V (2010) Bandwidth/fault tolerance/contention aware application-specific NoC using PSO as a mapping generator. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, pp 247–252

Chapter 22
On-Line Image Search Application Using Fast and Robust Color Indexing and Multi-Thread Processing

Wichian Premchaisawadi and Anucha Tungkatsathan

Abstract The keyword-based images search engine like Google or Yahoo may return a large number of junk images which are irrelevant to the given keyword-based queries. In this paper, an interactive approach is developed to filter out the junk images from the keyword-based Yahoo image search results through Yahoo' Boss API. The framework of multi-threaded processing is proposed to incorporate an image analysis algorithm into the text-based image search engines. It enhances the capability of an application when downloading images, indexing, and comparing the similarity of retrieved images from diverse sources. We also propose an efficient color descriptor technique for image feature extraction, namely, Auto Color Correlogram and Correlation (ACCC) to improve the efficiency of image retrieval system and reduce the processing time. The experimental evaluations based on the coverage ratio measure show that our scheme significantly improves the retrieval performance over the existing image search engines.

22.1 Introduction

Most of the popular, commercial search engines, such as Google, Yahoo, and even the latest application, namely Bing, introduced by Microsoft, have achieved great success on exploiting the pure keyword features for the retrieval process of large-

W. Premchaisawadi (✉) · A. Tungkatsathan
Graduate School of Information Technology in Business, Siam University,
38 Petkasem Rd., Phasi-Charoen, Bangkok, Thailand
e-mail: wichain@siam.edu

A. Tungkatsathan
e-mail: aimdala@hotmail.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*,
Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_22,
© Springer Science+Business Media B.V. 2011

scale online image collections. Unfortunately, these image search engines are still unsatisfactory because of the relatively low precision rate and the appearance of large amounts of junk images [1]. One of several main reasons is that these engines don't use visual signature of the image for image indexing and retrieval. The indexing images, retrieval processes, and similarity measure among images principally take a long computation time so that they aren't suitable for real-time process optimization approach. There are many researchers who are trying to minimize computation time by applying distributed computing, for instance cluster computing to reduce the computational time [2–7]. Lu et al. presented a parallel technique to perform feature extraction and a similarity comparison of visual features, developed on cluster architecture. The experiments conducted show that a parallel computing technique can be applied that will significantly improve the performance of a retrieval system [2]. Kao, et al. proposed a cluster platform, which supports the implementation of retrieval approaches used in CBIR systems. Their paper introduces the basic principles of image retrieval with dynamic feature extraction using cluster platform architecture. The main focus is workload balancing across the cluster with a scheduling heuristic and execution performance measurements of the implemented prototype [3]. Ling and Ouyang proposed a parallel algorithm for semantic concept mapping, which adopts two-stage concept searching method. It increases the speed of computing the low-level feature extraction, latent semantic concept model searching and bridging relationship between image low-level feature and global sharable ontology [4]. Kao presents techniques for parallel multimedia retrieval by considering an image database as an example. The main idea is a distribution of the image data over a large number of nodes enables a parallel processing of the compute intensive operations for dynamic image retrieval. However, it is still a partitioning of the data and the applied strategies for workload balancing [5]. Although, cluster computing is popularly used in images retrieval approaches, it only attacks this problem at the macro level. Especially, to design a distributed algorithm and program it with cross-platform capability is difficult. In contrast, this paper is concerned with the micro level aspect of the problem by using multi-threading. Multi-threading is not the same as distributed processing. Distributed processing which is sometimes called parallel processing and multi-threading are both techniques used to achieve parallelism (and can be used in combination) [8].

Fortunately, with the increasing computational power of modern computers, some of the most time-consuming tasks in image indexing and retrieval are easily parallelized, so that the multi-core architecture in modern CPUs and multi-threaded processing may be exploited to speed up image processing tasks. Moreover, it is possible to incorporate an image analysis algorithm into the text-based image search engines such as Google, Yahoo, and Bing without degrading their response time significantly [9]. We also presents modify advanced algorithm, namely auto color correlogram and correlation (ACCC) [10] based on a color correlogram (CC) [11], for extracting and indexing low-level features of images. The framework of multi-threaded processing for an on-line CBIR application is

proposed. It enhances the capability of an application when downloading images and comparing the similarity of retrieved images from diverse sources.

Section 22.2 presents the framework of an on-line image retrieval system with multithreading. Section 22.3 discusses the proposed indexing technique in older to speed up image processing tasks. The experimental study is presented in Sect. 22.4 and concluding remarks are set out in Sect. 22.5.

22.2 The Proposed Framework of Multithreading for an On-Line CBIR System

Before introducing our framework of multi-threading for an on-line CBIR application, we will briefly examine the properties of the queries to be answered. We have developed a novel framework of real-time processing for an on-line CBIR application, using relevance images from Yahoo images. Our method uses the following major steps: (a) Yahoo Images is first used to obtain a large number of images that are returned for a given text-based query; (b) The users select a relevance image and a user's feedback is automatically collected to update the input query for image similarity characterization; (c) A multi-threaded processing method is used to manage and perform data parallelism or loop-level parallelism such as downloading images, extraction of visual features and computation of visual similarity measures; (d) If necessary, users can also change a keyword before selecting a relevance image for the query; (e) The updated queries are further used to adaptively create a new answer for the next set of returned images according to the users' personal preferences (see Fig. 22.1). In this section, a multi-threaded processing method is used to carry out parallel processing of multiple threads for a specific purpose. Multi-threading is a way to let programs do more than one thing at a time, implemented within a single program, and running on a single system. The number of threads should be considered and they must technically be assigned to the correct parts of the program in order to utilize the threads more efficiently. The development of functions, classes, and objects in the program should logically be designed as a sequence of steps. In this research, we firstly use the threads to improve the downloading speed for images from various sources according to the locations specified in the .xml file that are returned from Yahoo BOSS API [12]. Second, they increase the speed of computing the image feature extraction and similarity measure of feature vectors. The framework of multi-thread processing is presented in Fig. 22.2. The thread control and the tasks insight of a thread for retrieving images are presented in Figs. 22.3 and 22.4.

An image list control receives the .xml files that are returned from Yahoo BOSS API. The lists of URL can be obtained from the .xml files. They are further displayed and used for downloading images from the hosts. An image download module is designed to work in a multithreaded process for downloading images from diverse sources. It is controlled by an image search control module. The image search control module performs a very important function in the

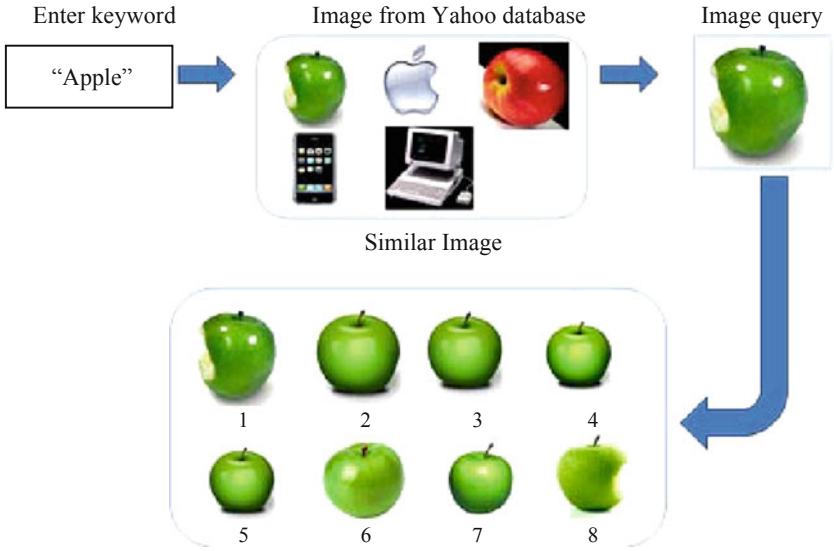


Fig. 22.1 Basic principles of the proposed system

management of the system. It fully supports and controls all modules of the online CBIR system. It checks for errors, and the input/output status of each module. Most importantly, it efficiently supports the synchronization of multiple threads that performs image download and similarity measurement by the associated modules. The similarity measurement module performs the computation of the feature vectors and distance metrics of all images that are obtained from the image download module. The image download and similarity measurement modules work concurrently. The query results are recorded into a session of an array in sequential order. The image list object is responsible for the arrangement of all displayed images on the application.

22.3 Feature Computation

This paper's main focus is on parallel computing techniques for image retrieval. The main objective is to reduce the processing time of real-time a CBIR system. However, an efficient color descriptor technique for image feature extraction is still required to reduce the processing time. In this section, we present an efficient algorithm for the proposed framework. It is a modifying of the correlogram technique for color indexing. An auto color correlation (ACC) [10] expresses how to compute the mean color of all pixels of color C_j at a distance kth from a pixel of color $C_j$in the image. Formally, the ACC of image$\{I(x, y),\ x = 1, 2, \ldots, M,\ y = 1, 2, \ldots, N\}$ is defined by Eq. 1.

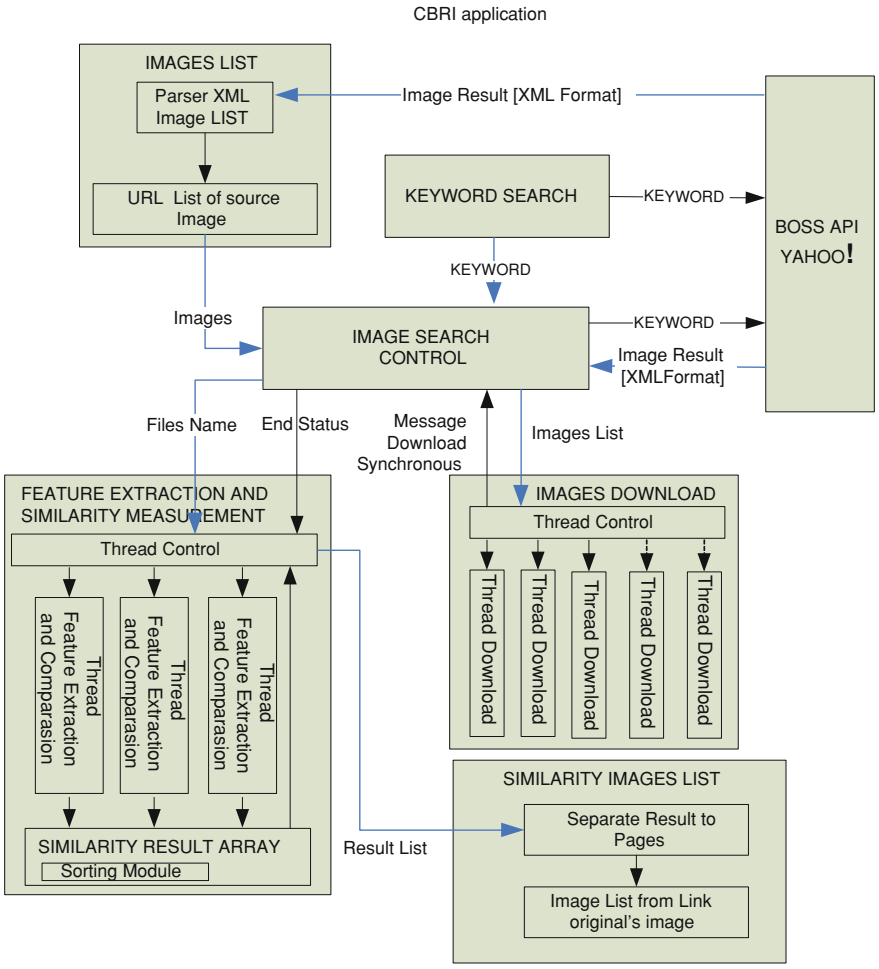


Fig. 22.2 The framework of in real-time multi-threaded processing for an on-line CBIR application

$$
\begin{aligned}
ACC(i,j,k) &= MC_j\gamma_{c_ic_j}^{(k)}(I) \\
&= \left\{ r_{mcj}\gamma_{c_ic_j}^{(k)}(I), g_{mcj}\gamma_{c_ic_j}^{(k)}(I), b_{mcj}\gamma_{c_ic_j}^{(k)}(I) | c_i \neq c_j \right\}
\end{aligned} \tag{22.1}
$$

where the original image $I(x,y)$ is quantized to m colors $C_1, C_2, \ldots, C_m$ and the distance between two pixels $d \in [\min\{M,N\}]$ is fixed a priori. Let MC_j is the color mean of the total number of color C_i from color C_i at distance kth in an image I. The arithmetic mean colors are computed by Eq. 22.2.

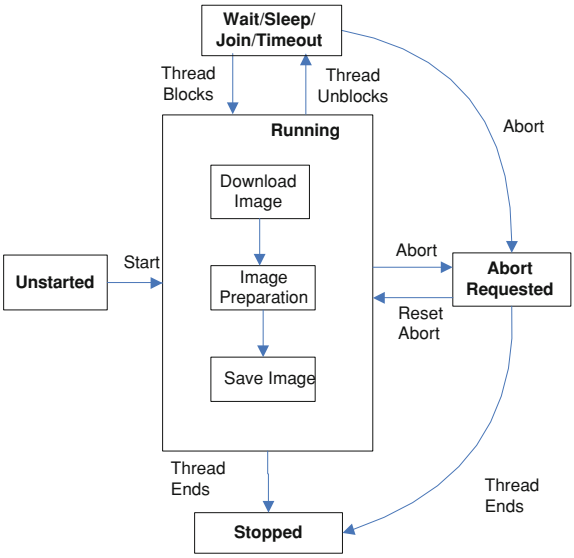


Fig. 22.3 The thread control and the tasks insight of the thread for downloading images

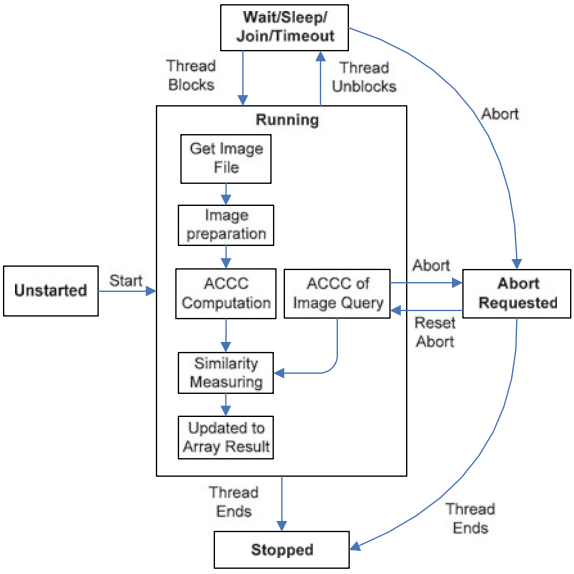


Fig. 22.4 The thread control and the tasks insight of the thread for retrieving images

$$
\begin{aligned}
r_{mc_j}\gamma_{c_ic_j}^{(k)}(I) &= \frac{\Gamma_{c_i,rc_j}^{(k)}(I)}{\Gamma_{c_i,c_j}^{(k)}(I)} | c_i \neq c_j \\
g_{mc_j}\gamma_{c_ic_j}^{(k)}(I) &= \frac{\Gamma_{c_i,gc_j}^{(k)}(I)}{\Gamma_{c_i,c_j}^{(k)}(I)} | c_i \neq c_j \\
b_{mc_j}\gamma_{c_ic_j}^{(k)}(I) &= \frac{\Gamma_{c_i,bc_j}^{(k)}(I)}{\Gamma_{c_i,c_j}^{(k)}(I)} | c_i \neq c_j
\end{aligned}
\tag{22.2}
$$

The denominator $\Gamma_{c_i,xc_j}^{(k)}(I)$ is the total of pixels color values of color C_j at distance k from any pixel of color C_i when xC_j is RGB color space of color C_j and denoted $C_j \neq 0$. N is the number of accounting color C_j from color C_i at distance k, defined by Eq. 22.3.

$$
N = \Gamma_{c_i,c_j}^{k}(I) = \left\{ \begin{array}{l} P(x_1, y_1) \in C_i | P(x_2, y_2) \in C_j; \\ k = \min\{|x_1 - x_2|, |y_1 - y_2|\} \end{array} \right\} \tag{22.3}
$$

We propose an extended technique of ACC based on the autocorrelogram, namely Auto Color Correlogram and Correlation (ACCC). It is the integration of Autocorrelogram [5] and Auto Color Correlation techniques [10]. However, the size of ACCC is still $O(md)$. The Auto Color Correlogram and Correlation is defined by Eq. 22.4.

$$
\text{ACCC}(j, j, k) = \left\{ \gamma_{c_i}^{(k)}(I), MC_j\gamma_{c_ic_j}^{(k)}(I) \right\} \tag{22.4}
$$

Let the ACCC pairs for the m color bin be (α_i, β_i) in I and (α_i', β_i') in I'. The similarity of the images is measured as the distance between the AC's and ACC's $d(I, I')$, which are derived from Lee et al. [13]. It is shown by Eq. 22.5

$$
d(I, I') = \left\{ \lambda_1 \sum_{\forall i} \frac{|\alpha_i - \alpha_i'|}{0.1 + \alpha_i + \alpha_i'} + \lambda_2 \sum_{\forall i} \frac{|\beta_i - \beta_i'|}{0.1 + \beta_i + \beta_i'} \right\} \tag{22.5}
$$

The λ_1 and λ_2 are the similarity weighting constants of autocorrelogram and auto color correlation, respectively. In the experiments conducted, $\lambda_1 = 0.5$ and α_1 and α_2 are defined by Eq. 22.6. The detail of ACC and ACCC algorithms are presented in Tungkastsathan and Premchaisawadi [10].

$$
\begin{aligned}
\alpha_i &= \gamma_{c_i}^{(k)}(I) \\
\beta_i &= \left\{ r_{mcj}\gamma_{c_ic_j}^{(k)}(I), g_{mcj}\gamma_{c_ic_j}^{(k)}(I), b_{mcj}\gamma_{c_ic_j}^{(k)}(I) | c_i \neq c_j \right\}
\end{aligned}
\tag{22.6}
$$

22.4 Experiment and Evaluation

The experiments that were performed are divided into two groups: In group 1, we evaluated the retrieval rate for on-line Yahoo image data sets in term of user relevance. And in group 3, we studied the performance of multi-thread processing in term of data parallelism for real-time image retrieval tasks.

22.4.1 Evaluated the Retrieval Rate

We have implemented an on-line image retrieval system using the Yahoo image database based on the Yahoo BOSS' API. The application is developed by using Microsoft .NET and implemented in the Windows NT environment. The goal of this experiment is to show that relevant images can be found after a small number of iterations, the first round is used in this experiment. From the viewpoint of user interface design, precision and recall measures are less appropriate for assessing an interactive system [14]. To evaluate the performance of the system in terms of user feedback, user-orientation measures are used. There have been other design factors proposed such as relative recall, recall effort, coverage ratio, and novelty ratio [15]. In this experiment the coverage ratio measure is used. Let R be the set of relevant images of query q and A be the answer set retrieved. Let $|U|$ be the number of relevant images which are known to the user, where $U \in R$. The coverage ratio is the intersection of the set A and U, $|R_k|$ be the number of images in this set. It is defined by Eq. 22.7.

$$\text{Coverage}(C_q) = \frac{|R_k|}{U} \tag{22.7}$$

Let $W_{(q)}$ is the number of keyword used. The average of coverage ratio is by Eq. 22.8.

$$\overline{C_{(q)}} = \frac{1}{N_{(q)}} \sum_{i=1}^{N_{(q)}} \frac{|R_k|}{|U|} \tag{22.8}$$

To conduct this experiment, Yahoo Images is first executed to obtain a large number of images returned by a given text-based query. The user selects a relevant image, specific to only one interaction with the user. Those images that are most similar to the new query image are returned. The retrieval performance in term of coverage ratio of the proposed system is compared to the traditional Yahoo text-based search results. The average coverage ratio is generated based on the ACC and ACCC algorithms using over 49 random test keywords in heterogeneous categories (i.e. animal, fruit, sunset, nature, and landscape). The results are presented in Table 22.1.

Table 22.1 Coverage ratio average of the top 24 of 200 retrieved images

Sample images	Coverage ratio				
	Animal	Fruit	Sunset/sunrise	Nature	Landscape
Sample 1	0.71	0.79	0.62	0.64	0.69
Sample 2	0.65	0.71	0.65	0.59	0.65
Avg.	0.68	0.75	0.63	0.62	0.67
Text-based	0.42	0.32	0.58	0.36	0.43

The data in a Table 22.1 shows that a user's feedback using a keyword with the ACCC algorithm can increase the efficiency of image retrieval from the Yahoo image database. Using the combination of text and a user's feedback for an image search, the images that do not correspond with the category are filtered out. It also decreases the opportunity of the images in other categories to be retrieved. In the experiment, we used two sample images obtained from the keyword search to test querying images for evaluating the performance of the system. The screenshots of the online image search application are shown in Figs. 22.5 and 22.6, respectively.

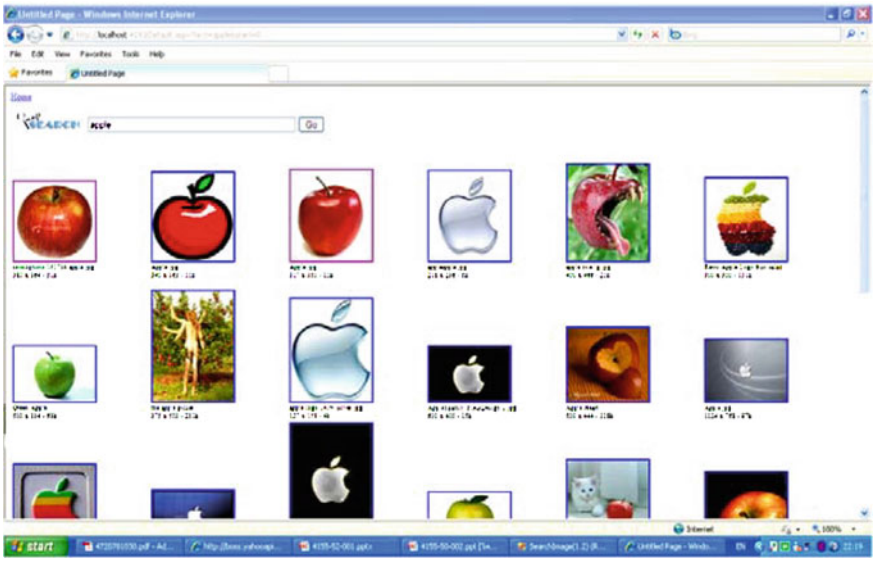


Fig. 22.5 Query results using a keyword search

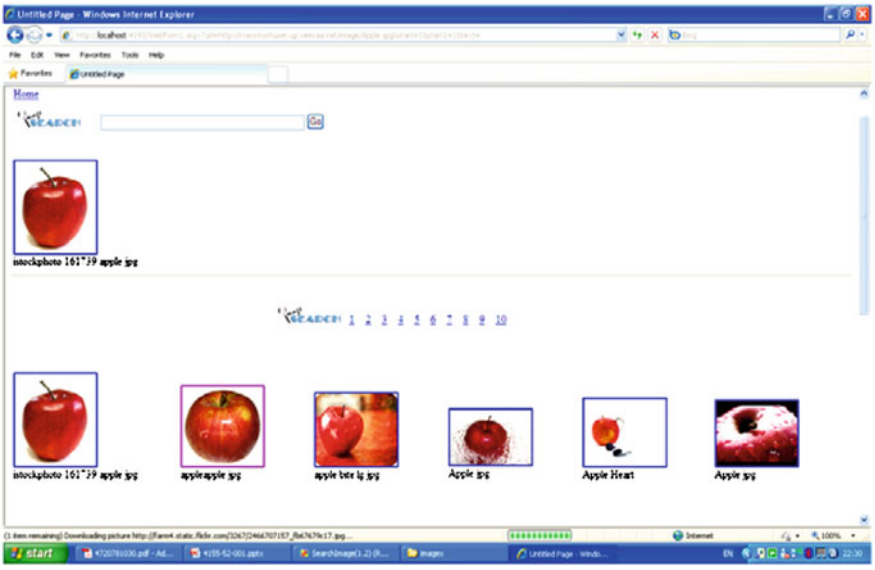


Fig. 22.6 Query results after applying a relevant feedback

22.4.2 Performance of Multithreading in the Image Retrieval Tasks

In the experimental settings, we used one keyword for downloading two hundred images and performed the image search in the same environment (internet speed, time for testing, hardware and software platforms). We tested the application by using 49 keywords in heterogeneous categories (i.e. animal, fruit, sunset, nature, and landscape). We tested the image search for three times in each keyword and calculated the average processing time of the whole process for an on-line image retrieval task. The number of downloaded images for each keyword had a maximum error value, which was less than ten percent of total downloaded images. The threads were tested and run on two different hardware platform specifications, single-core and multi-core CPUs.

The hardware specifications are described as follows. (1) Pentium IV single-core 1.8 GHz, and 1 GB RAM DDR2 system, (2) Quad-Core Intel Xeon processor E5310 1.60 GHz, 1066 MHz FSB 1 GB (2 × 512 MB) PC2-5300 DDR2. The number of threads versus time on single-core and multi-core CPUs for an image retrieval process that includes image downloading, feature extraction and image comparison, which are shown in our previous work [16]. We can conclude that the processing time for the same amount of threads in each platform for an image retrieval task is different (see in Figs. 22.7 and 22.8). However, we selected the most suitable number of threads from the tests on each platform to determine the assumptions underlying a hypothesis test. The results are shown in Table 22.2.

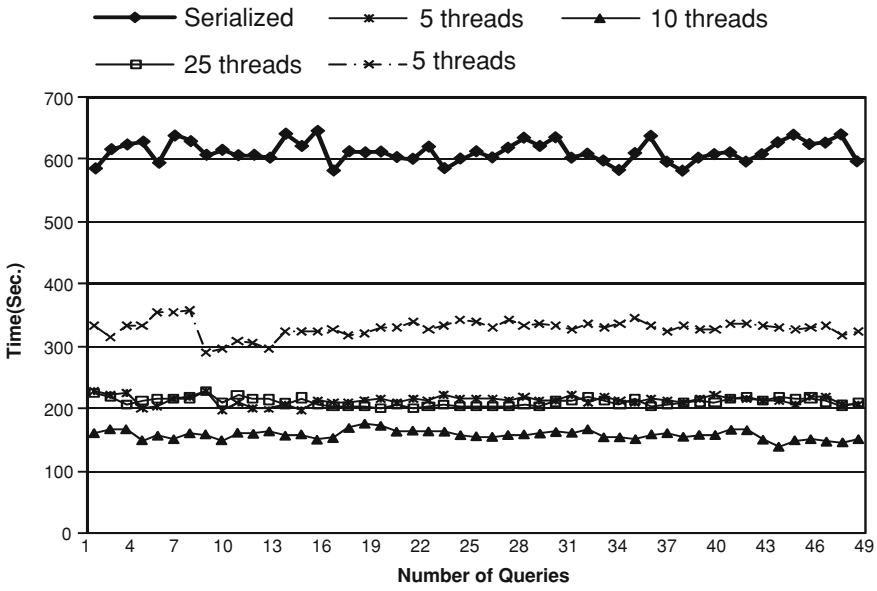


Fig. 22.7 Number of treads versus time on multi-core in all processes for online CBIR system [16]

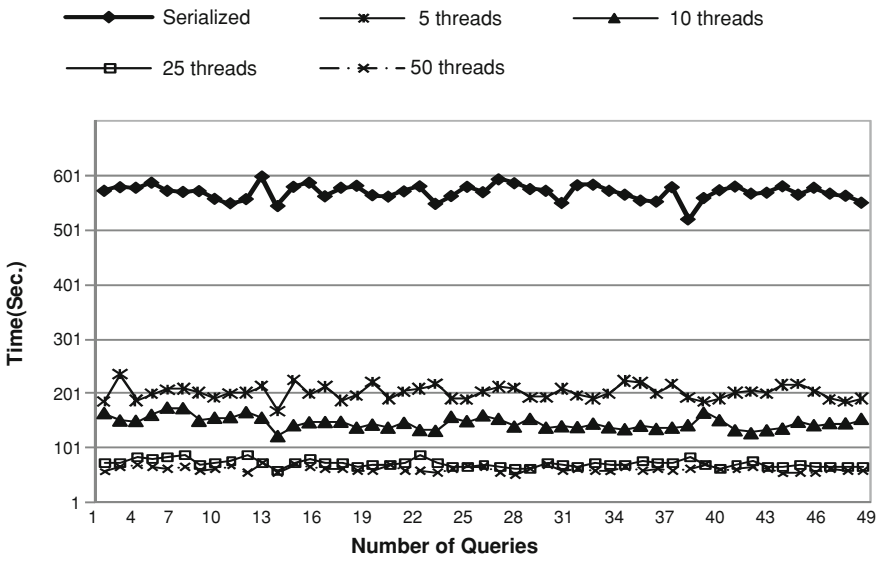


Fig. 22.8 Number of threads versus time on multi-core in all processes for online CBIR system [16]

Table 22.2 The average time in second of a whole process, image downloading, feature extraction, and image comparison at suitable number of threads in each platform (mean ± std-dev)

$W_{(q)}$	S-core 10 threads	Q-core 50 threads	$W_{(q)}$	S-core 10 threads	Q-core 50 threads
1	159.6 ± 3.3	57.0 ± 5.7	26	153.6 ± 9.2	61.0 ± 5.1
2	165.6 ± 13.5	57.3 ± 6.9	27	156.6 ± 4.9	64.3 ± 5.8
3	165.7 ± 15.1	65.7 ± 3.9	28	157.6 ± 13.1	68.3 ± 3.9
4	148.7 ± 18.9	70.0 ± 6.2	29	159.0 ± 11.4	66.0 ± 7.8
5	155.6 ± 4.9	75.3 ± 7.1	30	161.3 ± 11.8	56.7 ± 1.7
6	150.0 ± 13.1	53.3 ± 1.7	31	160.3 ± 7.6	51.3 ± 4.5
7	159.3 ± 8.5	60.7 ± 1.2	32	165.6 ± 16.9	63.3 ± 4.2
8	158.0 ± 21.4	65.7 ± 2.5	33	153.0 ± 7.5	69.3 ± 3.9
9	148.3 ± 17.6	61.3 ± 5.4	34	153.3 ± 12.6	59.0 ± 4.5
10	160.0 ± 21.9	67.0 ± 3.6	35	150.0 ± 10.7	63.6 ± 6.0
11	158.3 ± 5.7	71.0 ± 2.2	36	158.0 ± 11.3	63.7 ± 2.1
12	162.7 ± 10.9	66.3 ± 4.0	37	159.6 ± 11.4	60.7 ± 6.8
13	155.3 ± 3.4	63.7 ± 1.2	38	154.0 ± 5.9	61.0 ± 2.9
14	157.7 ± 3.1	62.0 ± 1.2	39	157.0 ± 13.4	65.7 ± 2.5
15	149.3 ± 4.5	58.7 ± 7.5	40	156.6 ± 8.9	62.3 ± 4.2
16	152.0 ± 2.3	61.3 ± 2.5	41	165.0 ± 11.1	53.7 ± 4.8
17	167.7 ± 18.6	66.0 ± 7.8	42	164.3 ± 9.7	56.7 ± 2.6
18	174.7 ± 15.1	66.0 ± 4.9	43	149.3 ± 11.1	56.3 ± 3.1
19	171.3 ± 7.6	58.0 ± 4.1	44	138.3 ± 6.2	64.3 ± 7.8
20	162.0 ± 10.0	71.0 ± 9.2	45	148.6 ± 4.0	58.0 ± 0.8
21	163.7 ± 3.7	59.3 ± 4.5	46	150.0 ± 9.9	57.7 ± 6.8
22	162.3 ± 8.3	58.0 ± 7.1	47	146.0 ± 7.5	57.0 ± 4.3
23	162.0 ± 4.2	56.0 ± 5.1	48	145.0 ± 8.5	60.7 ± 5.7
24	156.3 ± 16.0	72.3 ± 10.2	49	150.0 ± 10.4	54.0 ± 3.3
25	154.7 ± 8.9	64.3 ± 11.6	Avg	157.1 ± 8.4	62.0 ± 7.3

We formulated the hypothesis based on the experiment by using the statistical *t*-test. We did a *t*-test on the 49 keywords for retrieving images in order to measure the significance of the complete processing time obtained after applying our proposed scheme (see in Table 22.2). The mean processing times of single-core and multi-core platforms are 157.12 ± 8.4 and 62.0 ± 7.3, respectively. Using the *t*-test to compare the means of two independent CPU platform specifications, the *P* values obtained from the *t*-test of single-core versus multi-core is 1.98e−25. A statistical test shows that a multi-core platform significantly consumes less processing time than that of the single-core platform.

22.5 Conclusions

This research presents an interactive approach to filter out the junk images from the keyword-based Yahoo image search results. The advanced spatial color descriptors, namely; auto color correlation (ACC) and auto color correlogram

and correlation (ACCC), are proposed. In order for the processing time of feature computation to be reduced, the multi-threaded processing method is also proposed. The coverage ratio measure is used to evaluate the retrieval performance of the user's relevance feedback. Experiments on diverse keyword-based queries from Yahoo Images search engine obtained very positive results. Additionally, the experimental results show that our proposed scheme can speed up of the processing time for feature extraction and image similarity measurement as well as images downloading from various hosts. The use of multiple threads can significantly improve the performance of image indexing and retrieval on both platforms. In the future work based on this study, the distributed processing and multi-threading will be considered in combination to achieve the parallelism.

References

1. Yuli G, Jinye P, Hangzai L, Keim DA, Jianping F (2009) An interactive approach for filtering out junk images from keyword based Google search results. IEEE Trans Circuits Syst Video Technol 19(12):1–15
2. Lu Y, Gao P, Lv R, Su Z, Yu W (2007) Study of content-based image retrieval using parallel computing technique. In: Proceedings of the 2007 Asian technology information program's (ATIP's), 11 November–16 November 2007, China, pp 186–191
3. Kao O, Steinert G, Drews F (2001) Scheduling aspects for image retrieval in cluster-based image databases. In: Proceedings of first IEEE/ACM. Cluster computing and the grid, 15 May–18 May 2001, Brisbane, Australia, pp 329–336
4. Ling Y, Ouyang Y (2008) Image semantic information retrieval based on parallel computing. In: Proceeding of international colloquium on computing, communication, control, and management, CCCM, 3 August–4 August 2008, vol 1, pp 255–259
5. Kao O (2001) Parallel and distributed methods for image retrieval with dynamic feature extraction on cluster architectures. In: Proceedings of 12th international workshop on database and expert systems applications, Munich, Germany, 3 September 2001–7 September 2001, pp 110–114
6. Pengdong G, Yongquan L, Chu Q, Nan L, Wenhua Y, Rui L (2008) Performance comparison between color and spatial segmentation for image retrieval and its parallel system implementation. In: Proceedings of the international symposium on computer science and computational technology, ISCSCT 2008, 20 December–22 December 2008, Shanghai, China, pp 539–543
7. Town C, Harrison K (2010) Large-scale grid computing for content-based image retrieval. Aslib Proc 62(4/5):438–446
8. Multi-threading in IDL. http://www.ittvis.com/
9. Gao Y, Fan J, Luo H, Satoh S (2008) A novel approach for filtering junk images from Google search results. In: Lecture notes in computer science: advances in multimedia modeling, vol 4903, pp 1–12
10. Tungkastsathan A, Premchaisawadi W (2009) Spatial color indexing using ACC algorithms. In: Proceeding of the international conference on ICT and knowledge engineering, 1 December–2 December 2009, Bangkok, Thailand, pp 113–117
11. Huang J, Kumar SR, Mitra M, Zhu W-J (1998) Spatial color indexing and applications. In: Proceeding of sixth international conference on computer vision, 4 January–7 January 1998, Bombay, India, pp 606–607
12. Yahoo BOSS API. http://developer.yahoo.com/search/boss/

13. Lee HY, Lee HK, Ha HY, Senior member, IEEE (2003) Spatial color descriptor for image retrieval and video segmentation. IEEE Trans Multimed 5(3):358–367
14. Ricardo B-Y, Berthier R-N (1999) Modern information retrieval. ACM Press Book, New York
15. Robert RK (1993) Information storage and retrieval. Wiley, New York
16. Premchaisawadi W, Tungkatsathan A (2010) Micro level attacks in real-time image processing for an on-line CBIR system. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, pp 182–186

Chapter 23
Topological Mapping Using Vision and a Sparse Distributed Memory

Mateus Mendes, A. Paulo Coimbra and Manuel M. Crisóstomo

Abstract Navigation based on visual memories is very common among humans. However, planning long trips requires a more sophisticated representation of the environment, such as a topological map, where connections between paths are easily noted. The present approach is a system that learns paths by storing sequences of images and image information in a sparse distributed memory (SDM). Connections between paths are detected by exploring similarities in the images, using the same SDM, and a topological representation of the paths is created. The robot is then able to plan paths and switch from one path to another at the connection points. The system was tested under reconstitutions of country and urban environments, and it was able to successfully map, plan paths and navigate autonomously.

23.1 Introduction

About 80% of all the information humans rely on is visual [4], and the brain operates mostly with sequences of images [5]. View sequence based navigation is also extremely attractive for autonomous robots, for the hardware is very

M. Mendes (✉)
ESTGOH, Polytechnic Institute of Coimbra, R. General Santos Costa, 3400-124 Oliveira do Hospital, Portugal
e-mail: mmendes@estgoh.ipc.pt

M. Mendes · A. P. Coimbra · M. M. Crisóstomo
Institute of Systems and Robotics, Pólo II, University of Coimbra, 3000, Coimbra, Portugal
e-mail: acoimbra@deec.uc.pt

M. M. Crisóstomo
e-mail: mcris@isr.uc.pt

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_23,
© Springer Science+Business Media B.V. 2011

straightforward, and the approach is biologically plausible. However, while humans are able to navigate quite well based only on visual information, images usually require huge computer processing power. This means that for real time robot operation, visual information is often avoided. Other sensors, such as sonar or laser range finders, provide accurate information at a much lower computational cost.

The goal of equipping robots with cameras and vision-based navigation is still an open research issue. The use of special landmarks (possibly artificial, such as barcodes or data matrices) is a trick that can greatly improve the accuracy of the system [13]. As for the images, there are two popular approaches: one that uses plain images [6], the other that uses panoramic images [8]. Panoramic images offer a 360° view, which is richer than a plain front or rear view. However, that richness comes at the cost of even additional processing power requirements. Besides, the process of acquiring panoramic images requires the use of parabolic mirrors, which also introduce some distortion in the images. Some authors have also proposed techniques to speed up processing and/or reduce memory needs. Matsumoto [7] uses images as small as 32×32 pixels. Ishiguro [2] replaced the images by their Fourier transforms. Winters [15] compresses the images using Principal Component Analysis. All those techniques improve the processing time and/or efficiency of image processing in real time, contributing to make robot navigation based on image processing more plausible.

The images alone are a means for instantaneous localisation. View-based navigation is almost always based on the same idea: during a learning stage the robot learns a sequence of views and motor commands that, if followed with minimum drift, will lead it to a target location. By following the sequence of commands, possibly correcting the small drifts that may occur, the robot is later able to follow the learnt path on its own. The idea is very simple and it works very well for single paths. However, it is not versatile and requires that all the paths are taught one by one. For complex trips and environments, that may be very time consuming. The process can be greatly simplified using topological maps and path planning algorithms.

To plan paths efficiently, switching from one path to another at connection nodes, when necessary, more sophisticated representations of the environment are required than just plain images of sampling points. Those representations are provided by metric or topological maps [12]. Those maps represent paths and connections between them. They are suitable to use with search algorithms such as A*, for implementing intelligent planning and robot navigation.

This paper explains how vision-based navigation is achieved using a sparse distributed memory (SDM) to store sequences of images. The memory is also used to recognise overlaps of the paths and thus establish connection nodes where the robot can switch from one path to another. That way, a topological representation of the world can be constructed, and the system can plan paths. Part of this work has already been published in [9]. Section 23.2 explains navigation based on view sequences in more detail. Section 23.3 explains how the SDM works. In Sect. 23.4 the robot platform used for the experiments is described. Section 23.5 describes the navigation algorithm, and Sect. 23.6 shows and discusses the results obtained.

23.2 Navigation Using View Sequences

Usually, the vision-based approaches for robot navigation are based on the concept of a "view-sequence" and a look-up table of motor commands, where each view is associated with a corresponding motor command that leads the robot towards the next view in the sequence. In the present work, the approach followed is very similar to that of Matsumoto et al. [7]. That approach requires a learning stage, during which the robot must be manually guided. While being guided, the robot memorises a sequence of views automatically. While autonomously running, the robot performs automatic image based localisation and obstacle detection, taking action in real-time.

Localisation is estimated based on the similarity of two views: one stored during the learning stage and another grabbed in real-time. The robot tries to find matching areas between those two images, and calculates the horizontal distance between them in order to infer *how far* it is from the correct path. That distance is then used to correct possible drifts to the left or to the right. The technique is described in more detail in [10].

23.3 Sparse Distributed Memories

The sparse distributed memory is an associative memory model proposed by Kanerva in the 1980s [5]. It is suitable to work with high dimensional binary vectors. In the present work, an image can be regarded as a high-dimensional vector, and the SDM can be used simultaneously as a sophisticated storage and retrieval mechanism and a pattern-matching tool.

23.3.1 The Original Model

The underlying idea behind the SDM is the mapping of a huge binary memory onto a smaller set of physical locations, called *hard locations*. As a general guideline, those hard locations should be uniformely distributed in the virtual space, to *mimic* the existence of the larger virtual space as accurately as possible. Every datum is stored by distribution to a set of hard locations, and retrieved by *averaging* those locations and comparing the result to a given threshold. Figure 23.1 shows a model of a SDM. "Address" is the reference address where the datum is to be stored or read from. It will activate all the hard locations within a given access radius, which is predefined. Kanerva proposes that the Hamming distance, that is the number of bits in which two binary vectors are different, be used as the measure of distance between the addresses. All the locations that differ less than a predefined number of bits from the input address are selected for the read or write operation. In the figure,

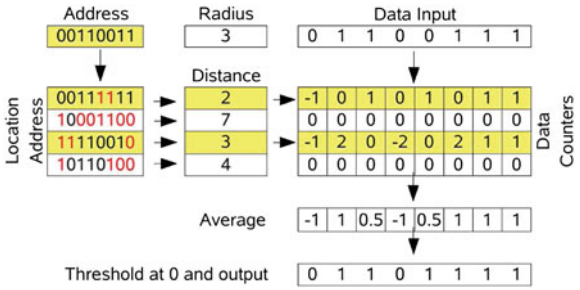


Fig. 23.1 One model of a SDM, using bit counters

the first and the third locations are selected. They dist, respectively, 2 and 3 bits from the input address, and the activation radius is exactly 3 bits.

Data are stored in arrays of counters, one counter for every bit of every location. Writing is done by incrementing or decrementing the bit counters at the selected addresses. To store 0 at a given position, the corresponding counter is decremented. To store 1, it is incremented. Reading is done by averaging the values of all the counters columnwise and thresholding at a predefined value. If the value of the sum is below the threshold, the bit is zero, otherwise it is one.

Initially, all the bit counters must be set to zero, for the memory stores no data. The bits of the address locations should be set randomly, so that the addresses would be uniformely distributed in the addressing space. There is no guarantee that the data retrieved is exactly the same that was written. It should be, providing that the hard locations are correctly distributed over the binary space and the memory has not reached saturation.

23.3.2 The Model Used

The original SDM model, though theoretically sound and attractive, has some faults. One problem is that of selecting the hard locations at random in the beginning of the operation. Another problem is that of using bit counters, which cause a very low storage rate of about 0.1 bits per bit of traditional computer memory and slow down the system. Those problems have been thoroughly described in [11], where the authors study alternative architectures and methods of encoding the data.

To overcome the problem of placing hard locations in the address space, in the present work the hard locations are selected using the Randomised Reallocation algorithm proposed by Ratitch and Precup [14]. The idea is that the system starts with an empty memory and allocates new hard locations when there is a new datum which cannot be stored in enough existing locations. The new locations are placed *randomly* in the neighbourhood of the new datum address. To overcome the problem of using bit counters, the bits are grouped as integers, as shown in Fig. 23.2. Addressing is done using an arithmetic distance, instead of the

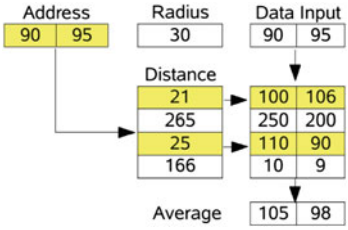


Fig. 23.2 Alternative architecture of the SDM, auto-associative and using integer numbers

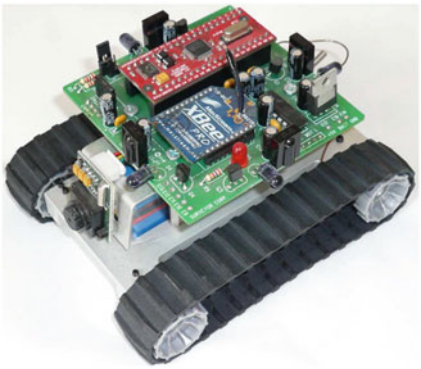


Fig. 23.3 Robot used

Hamming distance. Learning is achieved through the use of a gradient descent approach, updating each byte value using the equation:

$$h_t^k = h_{t-1}^k + \alpha \cdot (x^k - h_{t-1}^k), \quad \alpha \in \mathbb{R} \wedge 0 \leq \alpha \leq 1 \tag{23.1}$$

The value h_t^k is the kth integer number in the hard location h, at time t. The value x^k is the corresponding kth integer number in the input vector x. The coefficient α is the learning rate—in this case it was set to 1, enforcing one shot learning.

23.4 Experimental Platform

The robot used was a Surveyor SRV-1, a small robot with tank-style treads and differential drive via two precision DC gearmotors (Fig. 23.3). Among other features, it has a built in digital video camera and a 802.15.4 radio communication module. This robot was controlled in real time from a laptop with a 1.8 GHz processor and 1 Gb RAM. The overall software architecture is as shown in Fig. 23.4. It contains three basic modules:

1. The SDM, where the information is stored.
2. The Focus (following Kanerva's terminology), where the navigation algorithms are run.

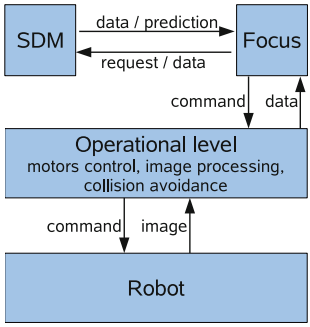


Fig. 23.4 Architecture of the implemented software

3. An operational layer, responsible for interfacing the hardware and some tasks such as motor control, collision avoidance and image equalisation. Navigation is based on vision, and has two modes: supervised learning, in which the robot is manually guided and captures images to store for future reference; and autonomous running, in which it uses previous knowledge to navigate autonomously, following any sequence previously learnt. The vectors stored in the SDM consist of arrays of bytes, as summarised in Eq. 23.2:

$$x_i = \langle im_i, seq\_id, i, timestamp, motion \rangle \tag{23.2}$$

In the vector, im_i is the image i, in PGM (Portable Gray Map) format and 80×64 resolution. In PGM images, every pixel is represented by an 8-bit integer. The value 0 corresponds to a black pixel, the value 255 represents a white pixel. *seq_id* is an auto-incremented, 4-byte integer, unique for each sequence. It is used to identify which sequence the vector belongs to. The number i is an auto-incremented, 4-byte integer, unique for every vector in the sequence, used to quickly identify every image in the sequence. The *timestamp* is a 4-byte integer, storing Unix timestamp. It is not being used so far for navigation purposes. The character *motion* is a single character, identifying the type of movement the robot performed after the image was grabbed.

The image alone uses 5,120 bytes. The overhead information comprises 13 additional bytes. Hence, the input vector contains a total of 5,133 bytes.

23.5 Mapping and Planning

The "teach and follow" approach per se is very simple and powerful. But for robust navigation and route planning, it is necessary to extend the basic algorithm to perform additional tasks. For example, it is necessary to detect connection points between the paths learnt, when two or more paths cross, come together or split apart. It is also necessary to disambiguate when there are similar images or divergent paths.

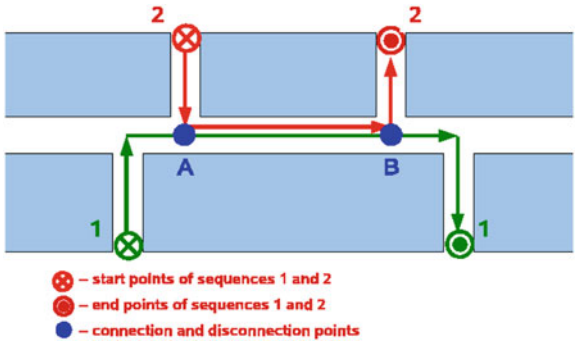


Fig. 23.5 Example of paths that have a common segment. The robot only needs to learn AB once

23.5.1 Filtering Out Unnecessary Images

During learning in vision-based navigation, not every single picture needs to be stored. There are scenarios, such as corridors, in which the views are very similar for a long period of time. Those images do not provide data useful for navigation. Therefore, they can be filtered out during the learning stage, so that only images which are *sufficiently* different from their predecessors must be stored. That behaviour can be easily implemented using the SDM: every new image is only stored if there is no image within a predefined radius in the SDM. If the error in similarity between the new image and any image in the SDM is below a given threshold, the new image is discarded. A good threshold to use for that purpose is the memory activation radius. Because of the way the SDM works, new images that are less than an activation radius from an already stored image will be stored in the same hard locations. Therefore, they are most probably unnecessary, and can be discarded with no risk of impairing the performance of the system.

23.5.2 Detecting Connection Points

Another situation in which new images do not provide useful information is the case when two paths have a common segment, such as depicted in Fig. 23.5. The figure shows two different paths, 1 and 2, in which the segment AB is common. If the robot learns segment AB for path 1, for example, then it does not need to learn it again for segment 2. When learning path number 2, it only needs to learn it until point A. Then it can store an association between paths 1 and 2 at point A and skip all the images until point B. At point B, it should again record a connection between paths 1 and 2. That way, it builds a map of the connection points between the known paths. That is a kind of topological representation of the environment.

The main problem with this approach is to detect the connection points. The points where the paths come together (point A in Fig. 23.5) can be detected after a

reasonable number of images of path 1 have been retrieved, when the robot is learning path 2. When that happens, the robot stores the connection in its working memory and stops learning path 2. From that point onwards, it keeps monitoring if it is following the same path that it has learnt. After a *reasonable* number of predictions have failed, it adds another connection point to the graph and resumes learning the new path. In the tests with the SDM, a number of 3–5 consecutive images within the access radius usually sufficed to establish a connection point, and 3–5 images out of the access radius was a good indicator that the paths were diverging again.

23.5.3 Sequence Disambiguation

One problem that arises when using navigation based on sequences is that of sequence disambiguation. Under normal circumstances, it is possible the occurrence of sequences such as (1) ABC; (2) XBZ; or (3) DEFEG, each capital letter representing a random input vector. There are two different problems with these three sequences: (1) and (2) both share one common element (B); and one element (E) occurs in two different positions of sequence (3). In the first case, the successor of B can be either C or Z. In the second case, the successor of E can be either F or G. The correct prediction depends on the history of the system. One possible solution relies on using a kind of *short term* memory.

Kanerva proposes a solution in which the input to the SDM is not the last input D_t, but the juxtaposition of the last k inputs $\{D_t, D_{t-1} \ldots D_{t-k}\}$. This technique is called *folding*, and k is the number of *folds*. The disadvantage is that it greatly increases the dimensionality of the input vector. Bose [1] uses an additional neural network, to store a measure of the *context*, instead of adding folds to the memory.

In the present work, it seemed more appropriate a solution inspired by Jaeckel and Karlsson's proposal of segmenting the addressing space [3]. Jaeckel and Karlsson propose to fix a certain number of coordinates when addressing, thus reducing the number of hard locations that can be selected. In the present work, the goal is to retrieve an image just within the sequence that is being followed. Hence, Jaeckel's idea is appropriate for that purpose. The number of the sequence can be *fixed*, thus truncating the addressing space.

23.6 Experiments and Results

For practical constraints, the experiments were performed in a small testbed in the laboratory. The testbed consisted of an arena surrounded by a realistic countryside scenario, or filled in with objects simulating a urban environment.

23.6.1 Tests in an Arena Stimulating a Country Environment

The first experiment performed consisted in analysing the behaviour of the navigation algorithm in the arena. The surrounding wall was printed with a composition of images of mountain views, as shown in Fig. 23.8. The field of view of the camera is relatively narrow (about 40°), so the robot cannot capture above or beyond the wall. Sometimes it can capture parts of the floor.

Figure 23.6 shows an example of the results obtained. In the example, the robot was first taught paths L1 and L2. Then the memory was loaded with both sequences, establishing connection points A and B. The minimum *overlapping* images required for establishing a connection point was set to 3 consecutive images. The minimum number of different images necessary for splitting the paths at point B was also set to 3 consecutive images out of the access radius. The lines in Fig. 23.6 were drawn by a pen attached to the rear of the robot. Therefore, they represent the motion of the rear, not the centre of the robot, causing the arcs that appear when the robot changes direction. As the picture shows, the robot was able to start at the beginning of sequence L1 and finish at the end of sequence L2, and vice versa. Regardless of its starting point, at point A it always defaulted to the only known path L1. This explains the small arc that appears at point A in path F2. The arc represents an adjustment of the heading when the robot defaulted to path L1.

The direction the robot takes at point B depends on the established goal. If the goal is to follow path L1, it continues along that path. If the goal is to follow path L2, it will disambiguate the predictions to retrieve only images from path L2. That behaviour explains the changes in direction that appear in the red line (F1) at point B. The arcs were drawn when the robot started at path L1, but with the goal of reaching the end of path L2.

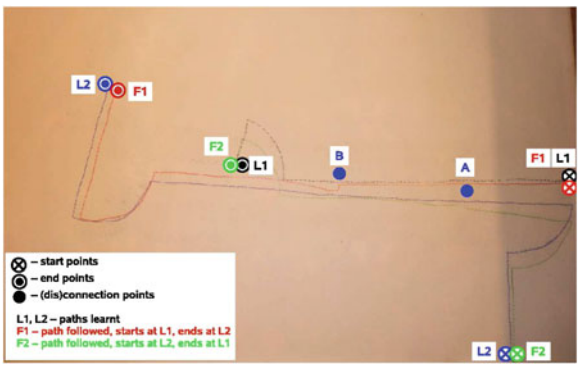
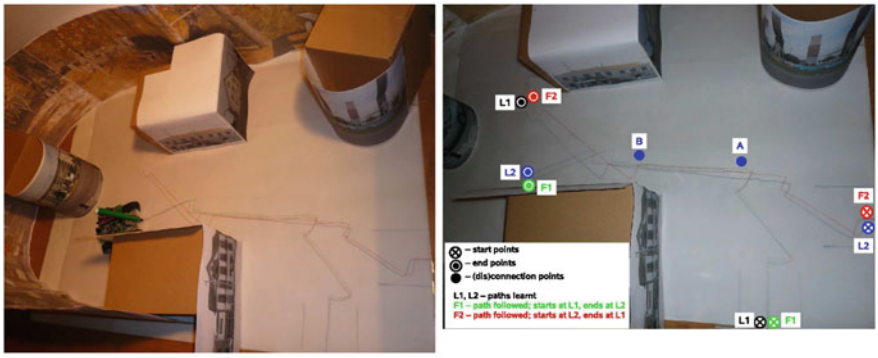


Fig. 23.6 Results: paths taught and followed. The robot successfully switches from one path to another and node points A and B

(a) (b)

Fig. 23.7 Typical city view, where the traffic turn is temporarily occluded by passing cars



(a) (b)

Fig. 23.8 Paths learnt (*blue* and *black*) and followed, with small scenario changes. The robot plans correctly the routes and is immune to small changes in the reconstituted urban scenario

23.6.2 Tests in a Stimulated Urban Environment

In a second experiment, the scenario was filled with images mimicking a typical city environment. Urban environments change very often. Ideally, the robot should learn one path in a urban environment but still be able to follow it in case there are small changes, up to an *acceptable* level. For example, Fig. 23.7 shows two pictures of a traffic turn, taken only a few seconds one after the other. Although the remaining scenario holds, one picture captures only the back of a car in background. The other picture captures a side view of another car in foreground.

Due to the small dimensions of the robot, it was not tested in a real city environment, but in a reconstruction of it. Figure 23.8 shows the results. Figure 23.8a shows the first scenario, where the robot was taught. In that scenario the robot, during segment AB, is guided essentially by the image of the traffic turn

without the car. In a second part of the same experiment, the picture of the traffic turn was replaced by the other picture with the car in foreground, and the robot was made to follow the same paths. Again, it had to start at path L1 and finish at path L2, and vice versa. As Fig. 23.8b shows, it was able to successfully complete the tasks.

23.7 Conclusions

Navigation based on view sequences is still an open research question. In this paper, a novel method was proposed that can provide vision-based navigation based on a SDM. During a learning stage, the robot learns new paths. Connection points are established when two paths come together or split apart. That way, a topological representation of the space is built, which confers on the robot the ability to switch from one sequence to another and plan new paths. One drawback of this approach is that the SDM model, simulated in software as in this case, requires a lot of processing and is not fast to operate in real time if the number of images is very large. Another disadvantage is that using just front views, the robot only merges paths that come together in the same heading. That problem can be solved using metric information to calculate when the robot is in a place it has already been, even if with another heading. Another possibility is to use omnidirectional images.

The results shown prove the feasibility of the approach. The robot was tested in two different environments: one that is a reconstitution of a country environment, the other a reconstitution of a changing urban environment. It was able to complete the tasks, even under changing conditions.

References

1. Bose J (2003) A scalable sparse distributed neural memory model. Master's thesis, University of Manchester, Faculty of Science and Engineering, Manchester, UK
2. Ishiguro H, Tsuji S (1996) Image-based memory of environment. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems
3. Jaeckel LA (1989) An alternative design for a sparse distributed memory. Technical report, Research Institute for Advanced Computer Science, NASA Ames Research Center
4. Johnson S (2004) Mind wide open. Scribner, New York
5. Kanerva P (1988) Sparse distributed memory. MIT Press, Cambridge
6. Matsumoto Y, Ikeda K, Inaba M, Inoue H (1999) Exploration and map acquisition for view-based navigation in corridor environment. In: Proceedings of the international conference on field and service robotics, pp 341–346
7. Matsumoto Y, Inaba M, Inoue H (2000) View-based approach to robot navigation. In: Proceedings of 2000 IEEE/RSJ international conference on intelligent robots and systems (IROS 2000)
8. Matsumoto Y, Inaba M, Inoue H (2003) View-based navigation using an omniview sequence in a corridor environment. In: Machine vision and applications

9. Mendes M, Paulo Coimbra A, Crisóstomo MM (2010) Path planning for robot navigation using view sequences. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, WCE 2010, London, UK
10. Mendes M, Crisóstomo MM, Paulo Coimbra A (2008) Robot navigation using a sparse distributed memory. In: Proceedings of the 2008 IEEE international conference on robotics and automation, Pasadena, CA, USA
11. Mendes M, Crisóstomo MM, Paulo Coimbra A (2009) Assessing a sparse distributed memory using different encoding methods. In: Proceedings of the 2009 international conference of computational intelligence and intelligent systems, London, UK
12. Meyer J (2003) Map-based navigation in mobile robots: Ii. A review of map-learning and path-planning strategies. Cogn Syst Res 4(4):283–317
13. Rasmussen C, Hager GD (1996) Robot navigation using image sequences. In: Proceedings of AAAI, pp 938–943
14. Ratitch B, Precup D (2004) Sparse distributed memories for on-line value-based reinforcement learning. In: ECML
15. Winters N, Santos-Victor J (1999) Mobile robot navigation using omni-directional vision. In: Proceedings of the 3rd Irish machine vision and image processing conference (IMVIP'99), pp 151–166

Chapter 24
A Novel Approach for Combining Genetic and Simulated Annealing Algorithms

Younis R. Elhaddad and Omar Sallabi

Abstract The Traveling Salesman Problem (TSP) is the most well-known NP-hard problem and is used as a test bed to check the efficacy of any combinatorial optimization methods. There are no polynomial time algorithms known that can solve it, since all known algorithms for NP-complete problems require time that is excessive to the problem size. One feature of Artificial Intelligence (AI) concerning problems is that it does not respond to algorithmic solutions. This creates the dependence on a heuristic search as an AI problem-solving technique. There are numerous examples of these techniques such as Genetic Algorithms (GA), Evolution Strategies (ES), Simulated Annealing (SA), Ant Colony Optimization (ACO), Particle Swarm Optimizers (PSO) and others, which can be used to solve large-scale optimization problems. But some of them are time consuming, while others could not find the optimal solution. Because of this many researchers thought of combining two or more algorithms in order to improve solutions quality and reduce execution time. In this work new operations and techniques are used to improve the performance of GA [1], and then combine this improved GA with SA to implement a hybrid algorithm (HGSAA) to solve TSP. This hybrid algorithm was tested using known instances from TSPLIB (library of sample instances for the TSP at the internet), and the results are compared against some recent related works. The comparison clearly shows that the HGSAA is effective in terms of results and time.

Y. R. Elhaddad (✉) · O. Sallabi
Faculty of Information Technology, Garyounis University, P.O. 1308, Benghazi, Libya
e-mail: younis_haddad@garyounis.edu

O. Sallabi
e-mail: Osallabi@garyounis.edu

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*,
Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_24,
© Springer Science+Business Media B.V. 2011

24.1 Introduction

Many problems of practical and theoretical importance within the fields of artificial intelligence and operations research are of a combinatorial nature. In these problems, there is a finite solution set $\boldsymbol{X}$ and a real-valued function $\boldsymbol{f}$: $\boldsymbol{X} \rightarrow \boldsymbol{R}$ whereby the goal is to search for a solution $\boldsymbol{x}^* \in \boldsymbol{X}$ with $\boldsymbol{f}(\boldsymbol{x}^*) \leq \boldsymbol{f}(\boldsymbol{x}) \; \forall \; \boldsymbol{x} \in \boldsymbol{X}$. The goal of an optimization problem can be formulated as follows: rearrange control or decision variables according to some constraints in order to minimize or maximize the value of an objective function [2]. The most widely known and famous example of a combinatorial optimization problem is the Traveling Salesman Problem (TSP) [2–4]. Problem-solving is an area of Artificial Intelligence (AI) that is concerned with finding or constructing the solution to a difficult problem like combinatorial optimization problems, using AI algorithms such as Genetic Algorithms (GA), Simulated Annealing (SA), Ant Colony Optimization (ACO), Particle Swarm Optimizers (PSO), Iterated Local Search (ILS), Tabu Search (TS), and others. These can be used to solve large-scale optimization problems. But some of them are time-consuming and others could not find the optimal solution because of the time constraints. Thus many researchers thought of combining two or more algorithms in order to improve solution quality and reduce execution time. In this work, new techniques and operations are applied to GA in order to improve its performance. Then this improved GA is combined with SA, using a new approach of this combination that produces a new Hybrid Genetic and Simulated Annealing Algorithm (HGSAA). The proposed algorithm was tested using symmetric TSP instances from known TSPLIB [5], and the results show that the algorithm is able to find an optimal solution or near optimal solution for varying sizes of these instances.

24.2 The Travelling Salesman Problem

Travelling Salesman Problem (TSP) is a classic case of a combinatorial optimization problem and is one of the most widely known Non deterministic Polynomial (NP-hard) problems [3]. The travelling salesman problem is stated as follows: given a number of cities with associated city to city distances, what is the shortest round trip tour that visits each city exactly once and then returns to the start city [6]. The TSP can be also stated as, given a complete graph, G, with a set of vertices, V, a set of edges, E, and a cost, c_{ij} associated with each edge in E, where c_{ij} is the cost incurred when traversing from vertex $i \in V$ to vertex $j \in V$, a solution to the TSP must return the minimum distance Hamiltonian cycle of G. A Hamiltonian cycle is a cycle that visits each node in a graph exactly once and returns to the starting node. This is referred to as a tour in TSP terms. The real problem is to decide in which order to visit the nodes. While easy to explain, this problem is not always easy to solve. There are no known polynomial time algorithms that can solve TSP. Therefore it is classified as an NP-hard problem. The TSP became

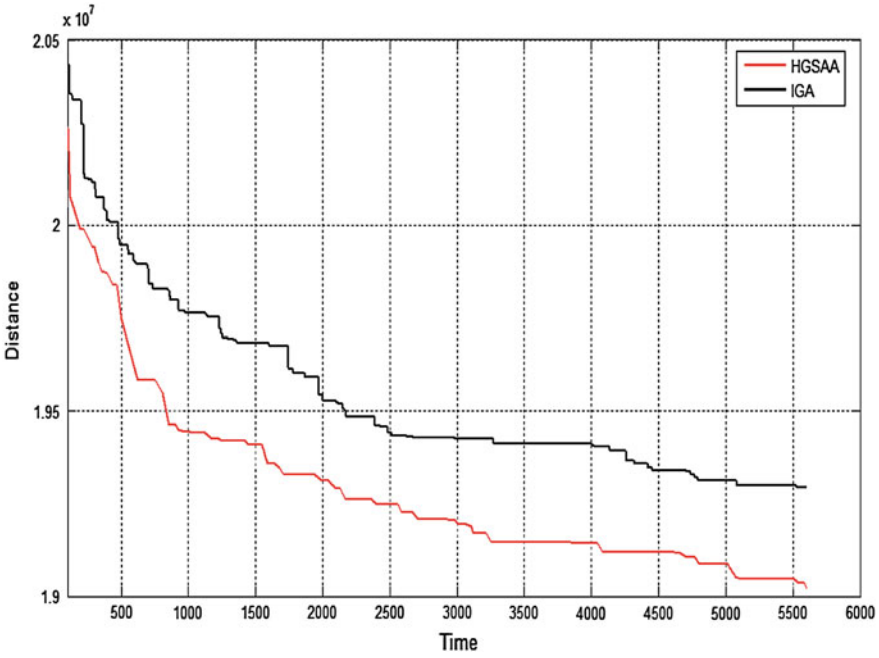


Fig. 24.1 Comparison of rate conversion for IGA and HGSAA

popular at the same time the new subject of linear programming arose along with challenges of solving combinatorial problems. The TSP expresses all the characteristics of combinatorial optimization, so it is used to check the efficacy of any combinatorial optimization method and is often the first problem researchers use to test a new optimization technique [7]. Different types of TSP can be identified by the properties of the cost matrix. The repository, TSPLIB, which is located at [5], contains many different types of TSP, and related problems. This thesis deals with symmetric TSP of type ECU_2D, where in symmetric (STSP) $c_{ij} = c_{ji}\ \forall i, j$, otherwise this set of problems is referred to as asymmetric (ATSP). The data of STSP given at TSPLIB contains the problem name (almost the name followed by the number of cities in the problem, e.g. kroA100, and rd100 both contain 100 cities in the problems). The data also provides the user with an array $n \times 3$ where n is the number of cities and the first column is the index of each city. Columns two and three are the positions of the city on the x-axis and the y-axis. Assuming that each city in a tour is marked by its position (x_i, y_i) in the plane (see Fig. 24.1), and the cost matrix c contains the Euclidean distances between the ith and jth city:

$$c_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{24.1}$$

The objective of TSP is to minimize the function f, where

$$f = \sum_{i=1}^{n-1} c_{i,i+1} + c_{1,n} \tag{24.2}$$

The search space of a Euclidean TSP of N cities contains $N!$ permutations. The objective is to find a permutation of the N cities that has minimum cost. For a symmetric problem with n cities there are $(n-1)!/2$ possible tours.

24.3 Genetic Algorithm

Evolutionary computation (EC) is based on biological evolution processes of living organisms, according to evolution theory of natural selection and survival of the fittest. EC consists of a population of individuals (solutions for a problem), performing iteratively. Operations such as reproduction, recombination, mutation and selection, result in the "survival of the fittest," or the best solution occurring in the population of solutions. Genetic algorithms (GAs) are a specific type of Evolutionary Algorithm (EA). GAs will be the center of attention appearing to be the best suited evolutionary algorithms for combinatorial optimization problems. The power of GAs comes from their reliable, robust optimization method and applicability to a variety of complex problems. In general GAs can be described as follows: Genetic algorithms start with generating random populations of possible solutions. Each individual of the population is represented (coded) by a DNA string, called a chromosome, and the chromosome contains a string of problem parameters. Individuals from the population are selected based on their fitness values. The selected parents are recombined to form a new generation. This process is repeated until some termination condition is met.

24.4 Simulated Annealing

The purpose of physical annealing is to accomplish a low energy state of a solid. This is achieved by melting the solid in a heat bath and gradually lowering the temperature in order to allow the particles of the solid to rearrange themselves in a crystalline lattice structure. This structure corresponds to a minimum energy state for the solid. The initial temperature of the annealing process is the point at which all particles of the solid are randomly arranged within the heat bath. At each temperature, the solid must reach what is known as thermal equilibrium before the cooling can continue [8]. If the temperature is reduced before thermal equilibrium is achieved, a defect will be frozen into the lattice structure and the resulting crystal will not correspond to a minimum energy state.

The Metropolis Monte Carlo simulation can be used to simulate the annealing method at a fixed temperature T. The Metropolis method randomly generates a sequence of states for the solid at the given temperature. A solid's state is characterized by the positions of its particles. A new state is generated by small movements of randomly chosen particles. The change in energy ΔE caused by the move is calculated and acceptance or rejection of the new state as the next state in the sequence is determined according to Metropolis acceptance condition. If $\Delta E < 0$ the move is acceptable and if $\Delta E > 0$ the move is acceptable with probability, if $e^{-\frac{-\Delta E}{t}} > \Omega$. The move is acceptable otherwise rejected, where Ω is random number and $0 < \Omega < 1$. Simulated annealing algorithms have been applied to solve numerous combinatorial optimization problems. The name and idea of SA comes from annealing in metallurgy, a technique involving heating and controlled cooling of a material to increase the size of its crystals and reduce their defects. The heat frees the atoms to move from their initial positions (initial energy). By slowly cooling the atoms the material continuously rearranges, moving toward a lower energy level. They gradually lose mobility due to the cooling, and as the temperature is reduced the atoms tend to crystallize into a solid. In the simulated annealing method, each solution s in the search space is equivalent to a state of a physical system and the function $f(s)$ to be minimized is equivalent to the internal energy of that state. The objective is to minimize the internal energy as much as possible. For successful annealing it is important to use a good annealing schedule, reducing the temperature gradually. The SA starts from a random solution x_p, selects a neighboring solution x_n and computes the difference in the objective function values, $\Delta f = f(x_n) - f(x_p)$. If the objective function is improved ($\Delta f < 0$), then the present solution x_p is replaced by the new one x_n; otherwise the solution that decreases the value of the objective function with a probability $pr = 1/(1 + e^{-\frac{\Delta E}{t}})$ is accepted, where pr is decreased as the algorithm progresses, and where (t) is the temperature or control parameter. This acceptance is achieved by generating a random number (rn) where $(0 \leq rn \leq 1)$ and comparing it against the threshold. If $pr > rn$ then the current solution is replaced by the new one. The procedure is repeated until a termination condition is satisfied.

24.5 Improved Genetic Algorithm Technique

Crossover is the most important operation of GA. This is because in this operation characteristics are exchanged between the individuals of the population. Accordingly (IGA) is concerned with this operation more than population size, thus the initial population consists of only two individuals, applying Population Reformulates Operation (PRO). Multi-crossovers are applied to these individuals to produce 100 children with different characteristics inherited from their parents, making ten copies of these children. Multi-mutation is applied, where each copy mutates with each method, evaluating the fitness function for each individual,

selects the best two individuals, and then finally applies the Partial Local Optimal (PLO) mutation operation to the next generation.

In the technique used for IGA the tour was divided into three parts with two randomly selected cut points (p_1 and p_2). The head contains $(1, 2, \ldots, p_1 - 1)$, the middle contains $(p_1, p_1 + 1, \ldots, p_2)$, and the tail contains $(p_2 + 1, p_2 + 2, \ldots, n)$. Using multi-crossover the head of the first parent is changed with the tail of the second parent. The middle remains unchanged, until partial local optimal mutation operation is applied which improves the middle tour by finding its local minima. The role of population reformulates operation is to change the structure of the tour by changing the head and the tail with the middle. In this technique the procedure ensures that new cities will be at the middle part of each cycle ready for improvement.

24.5.1 Multi-Crossover Operation

Crossover is the most important operation of GA because it exchanges characteristics between the individuals, and according to that many types of crossover operations are used to produce offspring with different attributes in order to build up an overall view of the search space. Multi-crossover works as mentioned below.

The basic principle of this crossover is two random cut points (p_1 and p_2), a head, containing $(1, 2, \ldots, p_1 - 1)$, the middle containing $(p_1, p_1 + 1, \ldots, p_2)$, and the tail containing $(p_2 + 1, p_2 + 2, \ldots, n)$. The head and tail of each parent are flipped, and then the head of the first parent is swapped with the tail of the other parent, and vice versa. For example, if the selected random two crossover points are $p_1 = 4$ and $p_2 = 7$, and two parents tours are:

$$\text{Parent1} \rightarrow \overbrace{9\ \ 1\ \ 5}^{\text{head1}}\ \overbrace{7\ \ 4\ \ 8\ \ 6}^{\text{mid1}}\ \overbrace{2\ \ 10\ \ 3}^{\text{tail2}}$$

$$\text{Parent2} \rightarrow \overbrace{2\ \ 8\ \ 5}^{\text{head2}}\ \overbrace{6\ \ 3\ \ 1\ \ 4}^{\text{mid2}}\ \overbrace{7\ \ 10\ \ 9}^{\text{tail1}}$$

For a valid tour the elements of head2 and tail2 are removed from the parent1 to give mid1

$$\overbrace{1\ \ 4\ \ 6\ \ 3}^{\text{mid1}}$$

In the same way, elements of head1 and tail1 are removed from the parent2 to give mid2

$$\overbrace{8\ \ 6\ \ 4\ \ 7}^{\text{mid2}}$$

Step 1 If the parts (head2, mid1, tail2) are reconnected using all possible permutations, six different children can be obtained (3!).

$$\text{child1} \rightarrow 2 \quad 8 \quad 5 \quad 1 \quad 4 \quad 6 \quad 3 \quad 7 \quad 10 \quad 9$$

In the same way for (head1, mid2, tail1), six other children are produced: i.e.

$$\text{child2} \rightarrow 9 \quad 1 \quad 5 \quad 8 \quad 6 \quad 4 \quad 7 \quad 2 \quad 10 \quad 3$$

Step 2 If the two heads are flipped, as in step 1, 12 new different children are produced:

$$\text{child3} \rightarrow 5 \quad 8 \quad 2 \quad 1 \quad 4 \quad 6 \quad 3 \quad 7 \quad 10 \quad 9$$

$$\text{child4} \rightarrow 5 \quad 1 \quad 9 \quad 8 \quad 6 \quad 4 \quad 7 \quad 2 \quad 10 \quad 3$$

Step 3 If the two tails are flipped and as in step 1, 12 new different children are produced:

$$\text{child5} \rightarrow 2 \quad 8 \quad 5 \quad 1 \quad 4 \quad 6 \quad 3 \quad 9 \quad 10 \quad 7$$

$$\text{child6} \rightarrow 9 \quad 1 \quad 5 \quad 8 \quad 6 \quad 4 \quad 7 \quad 3 \quad 10 \quad 2$$

Step 4 If the two mid are flipped and as in step 1; 12 new different children are produced:

$$\text{child7} \rightarrow 2 \quad 8 \quad 5 \quad 3 \quad 6 \quad 4 \quad 1 \quad 7 \quad 10 \quad 9$$

$$\text{child8} \rightarrow 9 \quad 1 \quad 5 \quad 7 \quad 4 \quad 6 \quad 8 \quad 2 \quad 10 \quad 3$$

Step 5 If the two heads and tails are flipped and as in step 1, 12 new different children are produced:

$$\text{child9} \rightarrow 5 \quad 8 \quad 2 \quad 1 \quad 4 \quad 6 \quad 3 \quad 9 \quad 10 \quad 7$$

$$\text{child10} \rightarrow 5 \quad 1 \quad 9 \quad 8 \quad 6 \quad 4 \quad 7 \quad 3 \quad 10 \quad 2$$

In each step 12 children are produced; therefore $5 \times (3!) \times 2 = 60$ completely different children are produced from just two parents.

24.5.2 Selection Operation

Using the rank selection, the best two individuals are selected for the next operations in order to reduce the execution time.

24.5.3 Mutation

The inversion mutation operation is used here, where random subtour is selected from the second individual then is inversed.

24.5.4 The Rearrangement Operation

This operation is applied to both individuals. $c_{i,j}$ is the cost between the two adjacent cities city_i and city_j, where $i = 1, 2, 3, \ldots, n-1$ and $j = i + 1$. The aim of this operation is to find the greatest (max) value of $c_{i,j}$ among all the adjacent cities on the tour, and then swap city_i with three other cities, one at a time. These cities are located on three different positions on the tour (beginning, middle, and end). The best position, as well as the original position will be accepted. This operation works in a random matter, and while it may not achieve any improvement after several iterations, it might instead (or is just as likely to) take a big jump and improve the result.

24.5.5 Partial Local Optimal Mutation Operation

In this operation, the subtour of individuals is selected randomly within the range of $3 \leq$ size of subtour $< n/4$. We then find the tour that produces the local minima of this subtour and exchange it with the original subtour. This operation is undertaken on one of the selected individuals after the mutation operation is performed.

24.6 The Proposed Hybrid Algorithm (HGSAA)

The proposed HGSAA is designed by combining the IGA and SA in order to reap the benefits of SA and reduce the time that IGA spends stuck at local minima. Initial temperature of SA is set at a small value, 80, because the number of cycles SA will perform is only ten cycles. Thus this temperature will ensure that SA can reach the state of equilibrium within these cycles. The hybrid algorithm starts with a random population. It will use the input of the GA, and multi-crossover is then applied to produce 60 different children. The parents' and their offspring's fitness will be calculated and depending on the results of this calculation a new population will be selected that is the same size as the original population. A partial local optimal mutation operation will then be applied to one individual (according to mutation probability) in order to improve its fitness value. The rearrangement operation is also used on the population. This process is continued until there is no improvement in the results after ten consecutive iterations. The memorized population from GA which provides the best result will then be transferred to the SA. The SA processes will be used to improve the results by using the nearest solution technique. If results are no longer improved within ten consecutive iterations, then the best memorized population from the SA will be moved to the GA to repeat the above process. Figure 24.1 shows the conversion rate of HGSAA and IGA for the

Table 24.1 Results of HGSAA

Problem	Optimal	Best result	Iteration	Time sec.	Average	St. dev.	Error (%)
eil101	629	**629**	400	17 (15)	632.9	2.8	0
ch130	6110	6126	500	26	6146.7	14.8	0.6
ch150	6528	**6528**	750 (**292**)	46 (**18**)	6540.4	13.9	0
korA100	21282	**21282**	400 (**171**)	18 (**7**)	21319.8	32.5	0
kroA150	26524	**26524**	800 (**407**)	53 (**27**)	26588.7	62.3	0
kroA200	29368	29382	1100	85	29434.9	45.7	0.23

dsj1000 problem from TSLIB [5]. Unlike the curve of HGSAA the curve of the IGA is stuck and the result is steady at many positions of the curve during its process. In other words in HGSAA the SA is causing the algorithm to be stuck for a long time and improves the results faster than the GA does.

24.7 Experimental Results of HGSAA

The following sections will discuss the results of experiments and compare them with some recently related work which used hybrid genetic algorithms to solve TSP.

24.7.1 Comparison with LSHGA

Instances that are ≥ 100 cities from TSPLIB [5] and used by Zhang and Tong [9] are used. The same number of generations for each instance is used in order to compare the results of HGSAA and the local search heuristic genetic algorithms LSHGA [9]. The HGSAA was run for ten trials corresponding to each instance, and the summarized results are shown in Table 24.1 where column 2 shows the known optimal solutions; column 3 shows the best result obtained by the HGSAA; column 4 indicates the number of generations performed, with the number of generations needed to obtain the optimal result in parentheses; column 5 indicates the time in seconds used for each instance, with the time to obtain the optimal result in parentheses; column 6 shows the average of the ten results for each instance; column 7 shows the standard deviation of the ten results for each instance; and column 8 shows the error ratio between the best result and the optimal, which is calculated according to Eq. 24.3. The results of LSHGA are summarized in Table 24.2 The notations, PS, CN, OS and error, denotes the population size of the algorithm, the convergence iteration number, the best solution of the LSHGA, and the error respectively. Errors are calculated according to Eq. 24.3.

Table 24.2 Results of LSHGA

Problem	PS	CN	BS	Error (%)
eil101	300	400	640	1.75
ch130	350	500	6164	0.88
ch150	400	750	6606	1.19
korA100	300	400	21296	0.66
kroA150	450	800	26775	0.95
kroA200	500	1100	29843	1.62

Table 24.3 Results of HGSAA and HGA

Instance	Optimal	HGSAA	HGA[35]
Eil51	426	426 (428.98)	426 (428.87)
Eil76	538	538 (544.37)	538 (544.37)
Eil101	629	629 **(640.2116)**	629 (640.975)*
KroA100	21282	21282	21282
KroD100	21294	21294	21306
D198	15780	15781	15788
kroA200	29368	29368	29368

$$\text{Error} = \frac{\text{average} - \text{optimal}}{\text{optimal}} \times 100. \tag{24.3}$$

From Tables 24.1 and 24.2 it is clear that the HGSAA performed better than the LSHGA. The HGSAA can find the optimal solution for four instances out of six, while LSHGA cannot find an optimal solution for any of the six instances. The error ratios in both tables indicate that the HGSAA performs much better than the LSHGA.

24.7.2 Comparison with HGA

The HGSAA has been compared to the HGA proposed by Andal Jayalakshmi et al. [10]. The HGSAA runs seven known instances of TSPLIB [5], ten trails for each one, same as the work done at [10]. The HGSAA used the integer and real tours eil51, eil76, and eil101. In Table 24.3 column 2 shows the known optimal solutions, column 3 shows the best result obtained by the HGSAA, the real number is in parenthesis; and column 4 indicates the best result of HGA from [10]. The comparison of the results summarized in table [3] shows that HGSAA obtained better results than the HGA. For real tours, for instance eil101, a new best result is obtained by HGSAA, where formerly the best known result was reported by [10].

Table 24.4 Results of HGSAA and SAGA

Problem	HGSAA		SAGA	
	Avg.	Std. dev.	Avg.	Std. dev.
dsj1000	**1.72**	**0.19**	2.27	0.39
d1291	**1.91**	**0.596**	3.12	1.12
fl1400	**0.43**	**0.38**	0.64	0.55
fl1577	**0.92**	**0.41**	0.64	0.55
pr2392	**6.37**	**0.36**	6.53	0.56

24.7.3 Comparison with SAGA

Stephen Chen and Gregory Pitt [11] proposed hybrid algorithms of SA and GA and they used large scale TSP. All of these instances were larger than 1,000 cities. Table 24.4 shows the average error from the optimal solution for each instance and the standard deviation of both HGSAA and SAGA. The termination condition for the HGSAA is set to be 7200 s for all except both fl1577 and pr2392 problems where the time for both is set to be 10800 s.

24.8 Conclusion and Future Work

As a scope of future work, possible directions can be summarized in the following points:

- To assess the proposed HGSAA, more empirical experiments may be needed for further evaluation of the algorithm. The announced comments may increase the effectiveness of the algorithm, thus should be discussed and taken into consideration.
- The data structures of the HGSAA algorithm can be refined. Therefore the execution time may be further reduced.
- Genetic Algorithms can be hybridized with another heuristic technique for further improvement of the results.
- The presented algorithm can be used to solve different combinational problems such as DNA sequencing.

References

1. Elhaddad Y, Sallabi O (2010) A new hybrid genetic and simulated annealing algorithm to solve the traveling salesman problem. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, vol I WCE 2010, June 30–July 2, London, UK, pp 11–14

2. Lawle EL (1976) Combinatorial optimization: networks and matroids. Holt, Rinehart, and Winston, New York
3. Larranaga P, Kuijpers CM, Murga RH, Inza I, Dizdarevic S (1999) Genetic algorithms for the travelling salesman problem: a review of representations and operators. *CiteSeerX*. citeseer.ist.psu.edu/318951.html. Accessed Nov 19, 2007
4. Fredman M et al (1995) Data structures for traveling salesmen. AT&T labs—research. www.research.att.com/~dsj/papers/DTSP.ps. Accessed Feb 13, 2008
5. Heidelberg University. http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/. Accessed Jan 22, 2007
6. Mitchell G, O'Donoghue D, Trenaman A (2000) A new operator for efficient evolutionary solutions to the travelling salesman problem. LANIA. www.lania.mx/~ccoello/mitchell00.ps.gz. Accessed Aug 22, 2007
7. Bhatia K (1994) Genetic algorithms and the traveling salesman problem. CiteSeer. http://citeseer.comp.nus.edu.sg/366188.html. Accessed Feb 26, 2008
8. Metropolis N et al (1953) Equation of state calculations by fast computing machines. Florida State University. www.csit.fsu.edu/~beerli/mcmc/metropolis-et-al-1953.pdf. Accessed Feb 17, 2008
9. Zhang J, Tong C (2008) Solving TSP with novel local search heuristic genetic algorithms. IEEE_explore. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4666929&isnumber=4666792. Accessed Jan 12, 2009
10. Jayalakshmi G, Sathiamoorthy S, Rajaram R (2001) A hybrid genetic algorithm—a new approach to solve traveling salesman problem. CiteSeer. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.3692. Accessed Jan 14, 2008
11. Chen S, Pitt G (2005) Isolating the benefits of respect. York University. http://www.atkinson.yorku.ca/~sychen/research/papers/GECCO-05_full.pdf. Accessed Jan 5, 2009

Chapter 25
Buyer Coalition Formation with Bundle of Items by Ant Colony Optimization

Anon Sukstrienwong

Abstract In electronic marketplaces, there are several buyer coalition schemes with the aim of obtaining the best discount and the total group's utility for buying a large volume of products. However, there are a few schemes focusing on a group buying with bundles of items. This paper presents an approach called GroupBuyACO for forming buyer coalition with bundle of items via the ant colony optimization (ACO). The concentration of the proposed algorithm is to find the best formation of the heterogeneous preference of buyers for earning the best discount from venders. The buyer coalition is formed concerning the bundles of items, item price, and the buyer reservations. The simulation of the proposed algorithm is evaluated and compared with the GAGroupBuyer scheme by Sukstrienwong (Buyer formation with bundle of items in e-marketplaces by genetic algorithm. Lecture note in engineering and computer science: proceedings of the international multiconference of engineers and computer scientists 2010, IMECS 2010, 17–19 March 2010, Hong Kong, pp 158–162). Experimental Results indicate that the algorithm can improve the total discount of any coalitions.

25.1 Introduction

At present, an electronic commerce is becoming a necessary tool for many companies to sell their products because it is one of the fastest ways to advertise the product's information to the huge number of customers. Tons of products can

A. Sukstrienwong (✉)
Information Technology Department, School of Science and Technology, Bangkok University, Bangkok, Thailand
e-mail: anon.su@bu.ac.th

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_25,
© Springer Science+Business Media B.V. 2011

be sold rapidly in few days. So, the companies can get better profits from selling a large number of products. Ordinarily, many sellers provide some attractive products with the special prices. One of the strategies which sellers prefer to make is selling their goods in bundles of item[1] with the special prices. Moreover, several commercial websites such as http://www.buy.yahoo.com.tw/ and https://www.shops.godaddy.com/ usually offer the volume discount for customers if the number of selling is big. For buyer side, most of the buyers prefer to build the corresponding purchasing strategies to reduce the amount of purchase cost. For this reason, the buyer strategy becoming rapidly popular on the Internet is a buyer coalition formation because buyers can improve their bargaining power and negotiate more advantageously with sellers to purchase goods at a lower price.

In the recent years, several existing buyer coalition schemes in electronic marketplaces have been developed. The main objective of these schemes is to gather all buyers' information for forming a buyer coalition to purchase goods at low cost. It helps to reduce the cost of communication and makes buyers comfortable in joining a coalition. The work of Ito et al. [10] presented an agent-mediated electronic market by group buy scheme. Buyers or sellers can sequentially enter into the market to make their decisions. The work of Tsvetovat et al. [18] has investigated the use of incentives to create buying group. Yamamoto and Sycara [20] presented the GroupBuyAction scheme for forming buyer coalition based on item categories. Then, the paper of Hyodo et al. [8] presented an optimal coalition formation among buyer agents based on genetic algorithms (GAs) with the purpose of distributing buyers among group-buying site optimally to get good utilities. The Combinatorial Coalition Formation scheme Li and Sycara [13] considers an e-marketplace where sellers set special offers based on volume. And, buyers place a bid on a combination of items with the reservation prices which is the maximum price that a buyer is willing to pay for an item of goods. In the work of Mahdi [14], GAs are applied for negotiating intelligent agents in electronic commerce using a simplified standard protocol. However, there are few schemes such as GroupPackageString scheme by Sukstrienwong [16] and GroupBuyPackage scheme by Laor et al. [11] that focused on a buyer coalition with bundles of items. Only the GroupPackageString scheme applied by using GAs to forms the buyer coalition with bundles of items.

In the corresponding conference paper, Sukstrienwong [17], to this paper, further results are found. The proposed approach applies ACO technique for forming buyer coalitions with the aim at maximizing the total discount. The paper is divided into five sections, including this introduction section. The rest of the paper is organized as follow. Section 25.2 outlines group buying with bundle of items and the motivating problem. Section 25.3 presents the basic concept of ACO and problem formulization to buyer formation with bundles of items. The experimental results of the simulation of the GroupBuyACO algorithm are in Sect. 25.4. The conclusions and future works are in last section.

[1] Bundle of items in the work of Gurler et al. [6] refers to the practice of selling two or more goods together in a package at a price which is below the sum of the independent prices.

Table 25.1 An example of price lists

Sellers	Package numbers	Product types				Price ($)
		Toilet paper	Paper tower	Lotion	Detergent	
s_0	$package_1^1$	pack of 1	–	–	–	8.9
	$package_2^1$	–	pack of 1	–	–	14.0
	$package_3^1$	–	pack of 3	–	–	32.5
	$package_4^1$	Pack of 1	Pack of 6	–	–	50.9
s_1	$package_1^2$	–	–	pack of 1		10.5
	$package_2^2$	–	–	–	pack of 1	19.0
	$package_3^2$	–	–	–	pack of 4	67.0
	$package_4^2$	–	–	pack of 1	pack of 8	92.0
s_3	$package_1^3$	–	pack of 1	–	–	14.0
	$package_2^3$	–	–	–	pack of 1	19.0
	$package_3^3$	–	pack of 3	–	pack of 1	49.5

25.2 Outline the Group Buying with Bundle of Items

In electronic marketplaces, sellers have more opportunity to sell their products in a large number if their websites are very well-known among buyers. Moreover, the pricing strategy is one of the tools for sellers that might expedite the selling volume. Some sellers simultaneously make a single take-it-or-leave-it price offer to each unassigned buyer and to each buyer group defined by Dana [2]. In this paper, I assume that the buyer group is formed under one goal to maximizing aggregate buyer's utility, the price discount received by being members of a coalition. Additionally, the definition of bundles of items is a slightly difference from the work of Gurler et al. [6]; in this paper, it refers to several items together in a package of one or more goods at one price. The discount policy of sellers based on the number of items bundled in the package. If the package is pure bundling, the average price of each item will be cheaper than the price of a single-item package. Suppose three sellers are in the e-marketplace selling some similar or the same products. Sellers prepare a large stock of goods and show the price list for each product. In this paper, I assume the agents are self-automate and be able to form coalitions when such a choice is beneficial. The example of three sellers' information is shown in Table 25.1. First seller, called s_1, is selling two sizes of facial toner, 100 and 200 cc. To get buyer attraction the seller s_1 has made the special offers. The Seller s_1 offers a package of number p_3^1 with the price of $32.0. The package p_3^1 composes of three bottle of facial toner (200 cc). The average price of each facial toner (200 cc) is about $32.0/3 = 10.67$ dollars/bottle which is $14.0 - 10.67 = 3.33$ dollars/bottle cheaper than a sing-bottle of facial toner (200 cc) in package p_2^1. At the same time, the third Seller called s_3 offers package p_3^3 which comprises of three bottles of facial toner (200 cc) and 1 bottle of body lotion (250 cc) at the price of $49.5. However, a single bottle of facial toner (200 cc) and body lotion (250 cc) are set individually in the package p_2^1 at the price

Table 25.2 An example of buyer's orders with the reservation price

bBuyers	Buyer's order (number of items reservation prices $)			
	Facial toner		Body lotion	
	100 cc	200 cc	1,500 cc	250 cc
$_1$	–	1 × (9.0)	–	1 × (10.5)
b_2	–	–	–	2 × (10.95)
b_3	–	–	3 × (6.0)	1 × (6.0)
b_4	1 × (8.0)	4 × (11.0)	–	–

of $14.0 and the package p_2^2 at the price of $19.0. Suppose there are some buyers who want to purchase some products listed in the Table 25.1. In the heterogeneous preference of buyers, some buyers do not want to purchase the whole bundle of items by their own. Buyers only need to buy a few items of products. Suppose a buyer called b_1 who wants to purchase a bottle of facial toner (200 cc) and a bottle of body lotion (250 cc) as shown in Table 25.2. Typically, buyers have seen the price lists provided by all sellers before making their orders. The problem of buyer b_1 is described as follows. If buyer b_1 goes straight to purchase those products by his own, the total cost that buyer b_1 needs to pay is $14.0 + 19.0 = 33.0$ dollars which is the highest price at that time. So, the buyer b_1 comes to participate in the group buying with the aim of obtaining better prices on the purchasing. Then, buyer b_1 places the orders to specific items with the reservation prices of $9.0 for facial toner (200 cc) and $10.5 a bottle of body lotion (250 cc).

25.3 Ant Colony Optimization for Buyer Coalition with Bundles of Items

25.3.1 The Basic Concept of ACO

The algorithm is based on an imitation of the foraging behavior of real ants as described in the work of Goss et al. [5]. Ant colony optimization (ACO) algorithms are inspired by the behavior of real ants for finding good solutions to combinatorial optimization. The first ACO algorithm was introduced by Dorigo and Gambardella [3] and Dorigo and Di Caro [4] which known as ant system (AS). ACO have applied to classical NP-hard combinatorial optimization problems, such as the traveling salesman problem in the work of Lawler et al. [12], the quadratic assignment problem (QAP) by Maniezzo et al. [15], the shop scheduling problem, and mixed shop scheduling by Yamada and Reeves [19]. The application of ACO appears in various fields. In the work of Ismail et al. [9], this paper presents the economic power dispatch problems solved using ACO technique. And, Alipour et al. [1] has proposed an algorithm based on ACO to enhance the quality of final fuzzy classification system.

In nature, real ants are capable of finding the shortest path from a food source to their nest without using visual cues shown by Hölldobler and Wilson [7]. In ACO, a number of artificial ants build solutions to an optimization problem while updating pheromone information on its visited tail. Each artificial ant builds a feasible solution by repeatedly applying a stochastic greedy rule. While constructing its tour, an ant deposits a substance called pheromone on the ground and follows the path by previously pheromone deposited by other ants. Once all ants have completed their tours, the ant which found the best solution deposits the amount of pheromone on the tour according to the pheromone trail update rule. The best solution found so far in the current iteration is used to update the pheromone information. The pheromone τ_{ij}, associated with the edge joining i and j, is updated as follow:

$$\tau_{ij} \leftarrow (1-\rho)\cdot\tau_{ij} + \sum_{k=1}^{m}\Delta\tau_{ij}^{k},, \tag{25.1}$$

where ρ is the evaporation rate which $\rho \in (0,1]$ the reason for this is that old pheromone should not have too strong an influence on the future. And $\Delta\tau_{ij}^{k}$ is the amount of pheromone laid on edge (i, j) by an ant k:

$$\Delta\tau_{ij}^{k} = \begin{cases} Q/L_k & \text{if edge}(i,\ j) \text{ is used by the ant } k \\ 0 & \text{otherwise,} \end{cases} \tag{25.2}$$

where Q is a constant, and L_k is the length of the tour performed by the ant k.

In constructing a solution, it starts from the starting city to visit an unvisited city. When being at the city i, the ant k selects the city j to visit through a stochastic mechanism with a probability p_{ij}^{k} given by:

$$p_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{c_{ij}\in N_i^k}\tau_{il}^{\alpha}\eta_{il}^{\beta}} & \text{if } j \in N_k^l \\ 0 & \text{otherwise,} \end{cases} \tag{25.3}$$

where N_k^i is a set of feasible neighborhood of ant k, representing the set of cities where the ant k has not been visited. α and β are two parameters which determine the relative influence of pheromone trail and heuristic information, and η_{ij}, which is given by

$$\eta_{ij} = \frac{1}{d_{ij}}, \tag{25.4}$$

where d_{ij} is the length of the tour performed by ant k between cities i and j.

25.3.2 Problem Formalization

There is a set of sellers on the Internet called $S = \{s_1, s_2, \ldots, s_m\}$ offering to sell a partial or all goods of $G = \{g_1, g_2, \ldots, g_j\}$. Let $B = \{b_1, b_2, \ldots, b_n\}$ denoted the

collection of buyers. Each buyer wants to purchase several items posted by some sellers in S. The seller i has made special offers within a set of packages, denoted as $\text{PACKAGE}_i = \{\text{package}_1^i,\ \text{package}_2^i, \ldots, \text{package}_k^i\}$. The average price of goods per item is a monotonically decreasing function when the size of the package is increasing big. A PACKAGE_i is associated with the set of prices, denoted $\text{PRICE}_i = \{\ \text{price}_1^i,\ \text{price}_2^i,\ \ldots,\ \text{price}_k^i\}$, where price_k^i is the price of package_k^i which is the combination of several items defined as $\text{package}_k^i = \{\text{g}_1^{i,k},\ \text{g}_2^{i,k}, \ldots,\ \text{g}_j^{i,k}\}$, $\text{g}_j^{j,k} \geq 0$. If any goods $\text{g}_j^{i,k}$ is not bundled in the package_k^i, then $\text{g}_j^{i,k} = 0$. Additionally, the product price of any seller, called s_m, is a function of purchased quantity, denoted $p_m(q)$, where q is the quantity of the product. The product price function is a monotonically decreasing function, $dp_m(q)/dq < 0$. If a buyer called b_m needs to buy some particular items offered by sellers in S, the buyer b_m places the order denoted as $Q_m = \{\text{q}_1^m,\ \text{q}_2^m, \ldots,\ \text{q}_j^m\}$, where q_j^m is the quantity of items g_j requested by the buyer b_m. If $\text{q}_j^m = 0$, it implies that the buyer b_m does have no request to purchase goods g_j. Additionally, the buyer b_m must put his reservation price for each goods associated with Q_m, denoted as $RS_m = \{rs_1^m,\ \text{rs}_2^m, \ldots,\ rs_j^m\}$ where $rs_h^m \geq 0,\ 0 \leq h \leq j$. In this paper, I assume that all of buyer reservation prices $\left(rs_h^m\right)$ of each item are higher than or equal to the minimum price sold by sellers. The objective of the problem is to find best utility of the coalition; the following terms and algorithm processes are needed to define. The coalition is a temporary alliance of buyers for a purpose of obtaining best utility. The utility of the buyer b_m gained from buying q_d^m items of g_d at the price_d is $\left(rs_d^m - \ \text{price}_d\right)q_d^m$. The total utility of the buyer b_m is

$$\sum_{d=1}^{j} (rs_d^m - \text{price}_d)q_d^m. \tag{25.5}$$

Then the total utility of the group is defined as follow:

$$U = \sum_{b_m \in B} \sum_{d=1}^{j} (rs_d^{b_m} - price_d)q_d^{b_m}, \tag{25.6}$$

where $j = |G|$.

25.3.3 Forming Buyer Group with Bundles of Items by Ants

The proposed algorithm presented in this paper provides means for buyer coalition formation by ACO. There some restrictions to this paper. Buyers are quoted a buyer-specific price after they have seen the price list of all packages provided by sellers. The buyer coalition is formed concerning only the price attribute. And, the price per item is a monotonically decreasing function when the size of the package

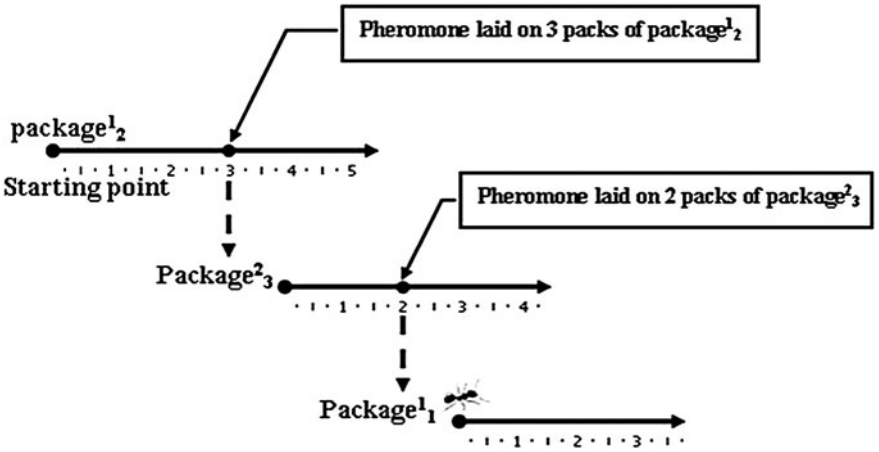


Fig. 25.1 Representing the work of one ant for creating the trail of <3 package$^1_2$2package$^2_3$…>

is increasing big. Additionally, the rule of the coalition is that each buyer is better forming a group than buying individually. The buyer coalition could not be formed if there is no utility earned from forming the group buyer.

The first step for forming buyer coalition with bundles of items is to represent the problem as a graph where the optimum solution is a certain way through this graph. In Fig. 25.1, the solid line represents a package selected by the ant k. If the selected package is picked more than one, the ant k moves longer along the solid line. Then, the ant k deposits and updates the pheromone on the selected number of the specific package. In this particular problem, the ant randomly chooses the other package which is represented by a dotted line. The probability of selecting i units of packages jth is p_{ij}^k formally defined below:

$$p_{ij}^k = \begin{cases} \frac{\tau_{ij}^{\alpha} n_{ij}^{\beta}}{\sum_i \sum_{1 \in D}^{j} \tau_{il}^{\alpha} n_{il}^{\beta}} & \text{if } j \in D, \text{ the set of packages offered by all sellers which have not been selected,} \\ 0 & \text{otherwise,} \end{cases} \tag{25.7}$$

where $\Delta\tau_{ij}^k$ is the intensity of the pheromone on the solid line. For instance, at the starting point if the ant k has selected three sets ($j = 3$) of package$^1_2$, the current ant deposits its pheromone only on the package$^1_2$, at the unit of 3. The ant k keeps moving along the path until all of the buyers' requests are matched. The possible resulting of the algorithm is shown in the Fig. 25.1. The quantity of pheromone $\Delta\tau_{ij}^k$ is defined as follow:

$$\Delta\tau_{ij}^k = \begin{cases} Q/U^k & \text{if } i \text{ units of package } j \text{ is used by the ant } k \\ 0 & \text{otherwise} \end{cases} \tag{25.8}$$

where Q is equal to one, and U^k is the total utility of a coalition derived from the ant k. Keep in mind that the η_{ij} is given by

$$\eta_{ij} = \begin{cases} \sum m_{ij}/u_{ij} & \text{if some items in the selected package are unmatched to the buyers' requests,} \\ 1 & \text{if all items in the selected package are totally matched to the buyers' request} \\ 0 & \text{otherwise} \end{cases} \tag{25.9}$$

where m_{ij} is the total number of items in the selected packages which is matched to the buyer's requests, and u_{ij} is the total number of items in the selected package.

At the beginning all of the pheromone values of each package line are initialized to the very small value c, $0 < c \leq 1$. After initializing the problem graph with a small amount of pheromones and defining each ant's starting point, a small number of ants run for a certain number of iterations. For every iteration, each ant determines a path through the graph from its starting point to the solid package line. The measurement of the quality of a solution found by the ACO is calculated according to the total utility of coalitions in Eq. 25.6.

25.3.4 GroupBuyACO Algorithm

```
1:  Procedure GroupBuyACO(){
2:  Initialization of the GroupBuyACO;
3:  Initialization  pheromone values to a small numerical constant c > 0
4:          while not (isFinish(Iteration )){
5:              for Ant = 1 to MaxAnt {
6:                  ManageAntsActivity();
7:                  EvaporatePheromone();
8:                  Save the best solution found so far.
9:                  UpdatePheromone();
10:             }
11:         }
12: }
13: ManageAntsActivity(){
14:         While not (isAntFinish(tour)){
15:             Select a package i to be visited and the exact j amount of
                package with probability p_ij^k in Eq. 25.7.
16:         }
17: }
18: EvaporatePheromone(){
    19:     Old pheromone should not have too strong an influence on
            the future. The evaporation rate value is ρ which is
            initialized to be small, ρ∈(0,1].
20: }
21: UpdatePheromone(){
22:         Update the all the path according to Eq. 25.8 and Eq. 25.9.
23: }
```

Table 25.3 Data settings for GroupBuyACO algorithm

Constant	Detail	Value
NumOfBuyer	No. of buyers	10
NumOfSeller	No. of sellers	3
MaxNumPackageSeller	Max no. of packages for each seller	5
NumOfTypeInPackage	No. of product type in pacakage	4

This section shows the implementation of ACO algorithm for forming buyer group with bundles of items called the GroupBuyACO algorithm. The proposed algorithm can be described by the following algorithm:

25.4 Experimental Results

This section demonstrates the initial data setting of the simulation for forming buyer coalition by the proposed algorithm, GroupBuyACO algorithm. The algorithm has tried several of runs with different numbers of artificial ants, values of α and β, and evaporation rate (ρ) to find which values would steer the algorithm towards the best solution.

25.4.1 Initial Data Settings

The experimental results of the proposed algorithm are derived from a simulation which has implemented more than 4,000 lines of C++ program. It is run on a Pentium(R) D CPU 2.80 GHz, 2 GB of RAM, IBM PC. The simulation program for the GroupBuyACO algorithm is coded in C++ programming language. Table 25.3 summarizes the initial data settings for GroupBuyACO algorithm in the simulation.

In order to get the best experimental results, for this example, the buyers' orders with the reservation price are selected randomly to demonstrate that the proposed algorithm is possible to works in the real-world data. Three different sellers offer to sell various packages which are pure bundling packages. Table 25.4 shows the products and price list offered by individual seller. Seller s_1 offers six packages. First four packages are one-item package. The rest are two-item package. The average number of items per package of s_1 is $(4 * 1 + 2 * 2)/6 = 1.33$. The seller s_2 combines two items of products in one package, so the average item per package is two. Seller s_3 has offered four packages of three items, so the average items per package for s_3 is three. And, there are ten buyers participating in the group buying shown in Table 25.5.

Table 25.4 The price list example for the simulation

Sellers	Package numbers	Product types				Price ($)
		A	B	C	D	
s_1	$package_1^1$	pack of 1	–	–	–	1,000
	$package_2^1$	–	pack of 1	–	–	1,000
	$package_3^1$	–	–	pack of 1	–	1,000
	$package_4^1$	–	–	–	pack of 1	1,000
	$package_5^1$	pack of 1	pack of 1	–	–	1,950
	$package_6^1$	–	pack of 1	pack of 1	–	1,900
s_2	$package_5^2$	–	–	pack of 1	pack of 1	1,925
	$package_6^2$	pack of 1	–	pack of 1	–	1,950
	$package_5^2$	pack of 1	–	–	pack of 1	1,920
	$package_6^2$	–	pack of 1	–	pack of 1	1,970
s_3	$package_1^3$	pack of 1	pack of 1	pack of 1	–	2,700
	$package_2^3$	–	pack of 1	pack of 1	pack of 1	2,690
	$package_3^3$	pack of 1	–	pack of 1	pack of 1	2,750
	$package_4^3$	pack of 1	pack of 1	–	pack of 1	2,700

Table 25.5 Buyer orders

bBuyers	Buyer's order (Number of items × (Reservation prices $))			
	A	B	C	D
b_1	–	–	1 × (970.0)	–
b_2	1 × (960.0)	1 × (975.0)	–	–
b_3	–	–	1 × (1000.0)	–
b_4	2 × (969.0)	–	–	–
b_5	–	1 × (955.0)	1 × (960.00)	–
b_6	–	–	–	1 × (980.00)
b_7	–	–	2 × (980.0)	–
b_8	–	4 × (970.0)	–	–
b_9	–	–	–	1 × (989.0)
b_{10}	1 × (965.0)	–	–	–

25.4.2 The GroupBuyACO Algorithm Performance

The first two parameters to be studies are α and β. As shown in Eq. 25.7, these parameters are related to the probability of selecting i units of packages jth (p_{ij}^k) because α is the exponent of $\Delta\tau_{ij}^k$ and β is the exponent of η_{ij}. Thus the corresponding variations in the values of both α and β might play an importance role on the GroupBuyACO algorithm. Let both α and β value range from 0.5 to 3, and the number of iterations is 200. The resulting of corresponding variation in the values of α and β is shown in Table 25.6. The best result is shown in bold. It can be seen

Table 25.6 The average of group's utility derived from corresponding in the values of α and β, iteration number = 2,000

α	β			
	0.5	1	2	3
0	759.06	457.22	791.33	673.21
0.5	755.25	594.23	814.71	734.72
1	623.57	757.48	698.21	542.01
2	927.24	907.09	456.98	671.45
3	554.65	657.84	569.24	459.27

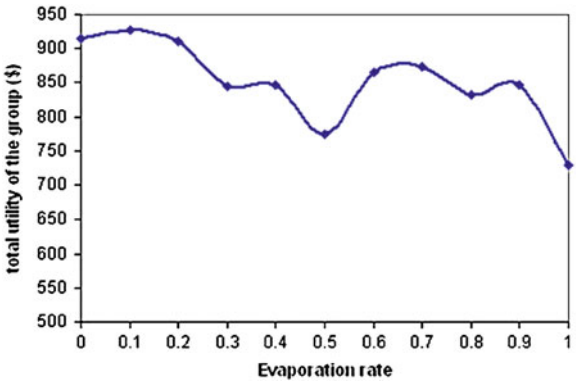


Fig. 25.2 Number of iterations where initial settings $\alpha = 2$, $\beta = 0.5$, and $\rho = 0.1$

Table 25.7 The comparison of GroupBuyACO algorithm with the genetic algorithm

GroupBuyACO algorithm ($)	GroupPackageString ($)
927.11	909.74

that the average utility of the group earned by GroupBuyACO algorithm was high when $\alpha = 2$ and $\beta = 0.5$.

Evaporation rate ρ of the pheromone is one of the most important variables for the GroupBuyACO algorithm. From Fig. 25.2, it can be seen that when the value of ρ is approximately 0.1, the total utility earned from the group buying is the highest. The proposed algorithm compared with the GAGroupBuyer scheme by Sukstrienwong [16]. In order to evaluate the performance of GroupBuyACO, the default configuration of parameters were set to the following values: $\alpha = 2$, $\beta = 0.5$ and $\rho = 0.1$. From Table 25.7, the GroupBuyACO algorithm outperforms GroupPackageString.

25.5 Conclusions and Future Work

In this paper, a new method for buyer coalition formation with bundle of items by ant colony optimization technique is proposed. The aim of the proposed algorithm is to form a buyer coalition in order to maximize the group's total utility. The ants

construct the trail by depositing pheromone after moving through a path and updating pheromone value associate with good or promising solutions through the edges of the path. From the experimental results, it is observed that the proposed algorithm is effective in dealing with finding best buyer coalitions with bundles of items. The solution quality of GroupBuyACO algorithm is shown by comparing with the genetic algorithm technique called GroupPackageString scheme. The experimental results show that the GroupBuyACO algorithm is able to yield better results than GAGroupBuyer scheme. However, the proposed algorithm has some restrictive constraints of forming a buyer coalition as follow: (1) all buyers quote specific prices for their requested products after they have seen the price list of all packages provided by sellers. (2) The buyer coalition is formed concerning only the price attribute. (3) And, the price per item is a monotonically decreasing function when the size of the package is increasing big. These restrictions can be extended to investigate in future researches.

References

1. Alipour H, Khosrowshahi Asl E, Esmaeili M, Nourhosseini M (2008) ACO-FCR: Applying ACO-based algorithms to induct FCR, Lecture note in engineering and computer science: proceedings of the world congress on engineering 2008, 2–4 July, London, UK, pp 12–17
2. Dana J (2004) Buyer groups as strategic commitments mimeo. Northwestern University, USA
3. Dorigo M, Gambardella LM (1997) Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans Evolut Comput 1(1):53–66
4. Dorigo M, Di Caro G (1999) The ant colony optimization metaheuristic. In: Corne D et al (eds) New ideas in optimization. McGraw Hill, London, pp 11–32
5. Goss S, Beckers R, Deneubourg JL, Aron S, Pasteels JM (1990) How trail laying and trail following can solve foraging problems for ant colonies. In: Hughes RN (ed) Behavioural mechanisms of food selection NATO-ASI Series, G 20. Springer, Berlin
6. Gurler U, Oztop S, Sen A (2009) Optimal bundle formation and pricing of two products with limited stock. J Int J Prod Econ,
7. Hölldobler B, Wilson EO (1990) The Ants. Springer, Berlin, p 732
8. Hyodo M, Matsuo T, Ito T (2003)An optimal coalition formation among buyer agents based on a genetic algorithm. In: 16th international conference on industrial and engineering applications of artificial intelligence and expert systems (IEA/AIE'03), Laughborough, UK, pp 759–767
9. Ismail M, Nur Hazima FI, Mohd. Rozely K, Muhammad Khayat I, Titik Khawa AR, Mohd Rafi A (2008) Ant colony optimization (ACO) technique in economic power dispatch problems. Lecture note in engineering and computer science: proceedings of the international multiconference of engineers and computer scientists, 19–21 March 2008, Hong Kong, pp 1387–1392
10. Ito T, Hiroyuki O, Toramatsu S (2002) A group buy protocal based on coalition formation for agent-mediated e-commerce. IJCIS 3(1):11–20
11. Laor B, Leung HF, Boonjing V, Dickson KW (2009) Forming buyer coalitions with bundles of items. In: Nguyen NT, Hakansson A, Hartung R, Howlett R, Jain LC (eds.) KES-AMSTA 2009. LNAI 5559-0717 Springer, Heidelberg, pp 121–138
12. Lawler EL, Lenstra JK, Rinnooy-Kan AHG, Shmoys DB (eds) (1985) The traveling salesman problem. Wiley, New York

13. Li C, Sycara K (2007) Algorithm for combinatorial coalition formation and payoff diversion in an electronic marketplace. In: Proceedings of the first international joint conference on autonomous agents and multiagent systems, pp 120–127
14. Mahdi S (2007) Negotiating agents in e-commerce based on a combined strategy using genetic algorithms as well as fuzzy fairness function. In: Proceedings of the world congress on engineering, WCE 2007, vol I. 2–4 July 2007, London, UK
15. Maniezzo V, Colorni A, Dorigo M (1994) The ant system applied to the quadratic assignment problem. Université Libre de Bruxelles, Belgium, Tech. Rep. IRIDIA/94-28
16. Sukstrienwong A (2010), Buyer formation with bundle of items in e-marketplaces by genetic algorithm. Lecture note in engineering and computer science: proceedings of the international multiconference of engineers and computer scientists 2010, IMECS 2010, 17–19 March 2010, Hong Kong, pp 158–162
17. Sukstrienwong A (2010) Ant colony optimization for buyer coalition with bundle of items. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July, London, UK, pp 38–43
18. Tsvetovat M, Sycara KP, Chen Y, Ying J (2001)Customer coalitions in electronic markets. Lecture notes in computer science, vol 2003. Springer, Heidelberg, pp 121–138
19. Yamada T, Reeves CR (1998) Solving the Csum permutation flowshop scheduling problem by genetic local search. In: Proceedings of 1998 ieee international conference on evolutionary computation, pp 230–234
20. Yamamoto J, Sycara K (2001) A stable and efficient buyer coalition formation scheme for e-marketplaces. In: Proceedings of the 5th international conference on autonomous agents, Monttreal, Quebec, Canada, pp 576–583

Chapter 26
Coevolutionary Grammatical Evolution for Building Trading Algorithms

Kamal Adamu and Steve Phelps

Abstract Advancements in communications and computer technology has enabled traders to program their trading strategies into computer programs (trading algorithms) that submit electronic orders to an exchange automatically. The work in this chapter entails the use of a coevolutionary algorithm based on grammatical evolution to produce trading algorithms. The trading algorithms developed are benchmarked against a publicly available trading system called the turtle trading system (TTS). The results suggest that out framework is capable of producing trading algorithms that outperform the TTS. In addition, a comparison between trading algorithms developed under a utilitarian framework, and using Sharpe ratio as objective function shows that they have statistically different performance.

26.1 Introduction

Traders make trade decisions specifying *entry*, *exit*, and *stop loss* prices [1]. The *entry* is the price at which the trader wishes to *enter* the market, the *exit* is the price at which the trader expects to take profit, and the *stop loss* is the price at which the trader wants to *exit* a position when a trade is not in her favour [1]. A set of *entry*, *exit*, and *stop loss* rules is referred to as a *trading system* and

K. Adamu (✉) · S. Phelps
Center for Computational Finance and Economic Agents, University of Essex, Colchester, UK
e-mail: ksadam@essex.ac.uk

S. Phelps
e-mail: sphelps@essex.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_26,
© Springer Science+Business Media B.V. 2011

there exists an interdependency between these rules [1]. A trader that consistently fails to *exit* a loosing trade when they have incurred a tolerable amount of loss will almost certainly be wiped out after a couple of loosing trades. Moreover, a trader that takes profit too early or too late before making a required amount of profit will have very little to cover their costs and loss or lose part of the profit she has made [1]. Technicians decide on *entry*, *exit*, and *stop loss* prices based on technical trading rules [1].

Advancements in communication and computer technology has allowed traders to submit trades electronically using computer programs (Trading algorithms) that sift through a vast amount of information looking for trade opportunities [2]. Trading algorithms have gained popularity due to their cost effective nature [2]. According to Hendershott et al. [2] 75% of trades executed in the US in 2009 were by trading algorithms. The aim of the work in this chapter is to test if a methodology based on grammatical evolution (GE) [3] can be used to coevolve rules for *entry*, *exit*, and *stop loss* that outperform a publicly available trading system called the turtle trading system in high frequency [1, 4]. This chapter also tests if trading algorithms developed under a utilitarian framework have the similar performance as trading algorithms developed using the Sharpe ratio as objective function.

Adamu and Phelps, Saks and Maringer [5, 6] employ cooperative coevolution in developing technical trading rules. In this chapter, we coevolve rules that form trading algorithms using GE for high frequency trading. The trading algorithms evolved are benchmarked against the turtle trading system. In addition, the effect of various objective functions on the trading algorithms evolved is considered.

The rest of this chapter is organised as follows. Section 26.2 gives a survey on the turtle trading system, and investor preference . We explain our framework in Sect. 26.3 and present the data used for the study in Sect. 26.4 Our result is presented in Sect. 26.5 and the Chapter ends with a summary in Sect. 26.6.

26.2 Background

26.2.1 The Turtle Trading System

A trading system is a set of rules that signal when to *enter*, and *exit* a position where a position is a steak in a particular asset in a particular market [1]. The rules in trading systems specify when to enter the market when prices are expected to fall, when to enter the market when prices are expected to rise, and how to minimise loss and maximise profit (Money management) [1]. The *entry* rules for the turtle trading system are specified as follows [1, 4]:

if $H_t > H_{t-55}$ or $H_t > H_{t-20}$ **then**
  GO LONG
else if $L_t < L_{t-55}$ or $L_t < L_{t-20}$ **then**
  GO SHORT
end if

and the *exit* rules for the turtle trading system are as follows:

if $L_t < H_{t-20}$ or $L_t < L_{t-10}$ **then**
  Exit long position
else if $H_t > L_{t-20}$ or $H_t > H_{t-10}$ **then**
  Exit short position
end if

$H_t, t \in \{1, 2, 3, 4, \ldots, T\}$ is the current highest price, and $L_t, t \in \{1, 2, 3, 4, \ldots, T\}$ is the current lowest price. The TTS places the initial *stop loss* at *entry* using the following equation:

$$Stop_t = \begin{cases} Stop_{t-1} - 2\text{ATR}_t & \text{if Long} \\ Stop_{t-1} + 2\text{ATR}_t & \text{if short.} \end{cases} \tag{26.1}$$

where ATR_t is the current average *true range* and it is calculated as follows:

$$\text{ATR}_t = \frac{19N_{t-1}\text{TR}_t}{20} \tag{26.2}$$

$\text{TR}_t, t \in \{1, 2, 3, 4, \ldots, T\}$ is the *true range* and its calculated as follows:

$$\text{TR}_t = \max(H_t - L_t,\ H_t - C_{t-1},\ C_{t-1} - L_t) \tag{26.3}$$

$C_t, t \in \{1, 2, 3, 4, \ldots, T\}$ is the price at the end of the time interval $t, t \in \{1, 2, 3, \ldots, T\}$.

26.2.2 Investor Preference

26.2.2.1 Utility Theory

Traditional finance and economics postulates that the financial markets are populated by rational, risk averse agents that prefer more wealth to less wealth [7]. One of the corner stones of efficient markets is the presence of the $homo - econimucus$, the rational risk averse agent with preference for more wealth than less wealth [6, 7] . In utilitarian terms this translates to expected utility maximising investors with utility functions that satisfy $U(W)^{'} > 0$ and $U(W)^{''} < 0$ where U(W) is the utility of wealth, and W is the current level of wealth [7]. The power utility function, negative exponential utility function, and quadratic utility function satisfy $U(W)^{'} > 0$ and $U(W)^{''} < 0$.

The *Power utility function* is defined by the following equation [7]

$$U(W) = \frac{W^{1-\gamma}}{1-\gamma}, \quad \gamma > 0, \quad \gamma \neq 1 \tag{26.4}$$

γ controls the degree of risk aversion of the utility function. There is evidence to suggest that traders exhibit loss aversion as well as risk aversion [6] hence in this chapter, a modified wealth is used to account for loss aversion [6]. The power utility function (PUF) is then defined as follows [6]:

$$U(W_i) = \frac{{W_i}^{1-\gamma}}{1-\gamma} - \frac{1}{1-\gamma}, \quad \gamma > 1 \tag{26.5}$$

$$W_i = \begin{cases} W_0(1+v_i) & v_i > 0 \\ w_0(1+v_i)^{\lambda} & v_i < 0, \lambda > 1 \end{cases} \tag{26.6}$$

where v_i is the simple return for trade interval $i, i \in \{1, 2, \ldots, N\}$; W_i is a modified level of wealth for the given trade interval $i, i \in \{1, 2, \ldots, N\}$. For this study we consider the case of a unit investor and set the initial level of wealth $W_0 = 1$. λ, and γ define the risk, and loss preference of the agents respectively.

Quadratic utility function. The quadratic utility function (QUF) is given by the following equation [7]:

$$U(W_i) = W_i - \frac{b}{2}W_i^2, \quad b > 0 \tag{26.7}$$

where W is the wealth. To satisfy the condition of $U(W)' > 0$, we set $W = \frac{1}{b}$ for levels of wealth $W > \frac{1}{b}$.

The *Negative exponential utility function* is given by the following equation [7]:

$$U(W_i) = a - be^{-cW_i}, \quad c > 0 \tag{26.8}$$

26.2.2.2 Sharpe Ratio

It follows that, provided a utility function satisfies $U(W)' > 0$ and $U(W)'' < 0$ then it suffices to look at the mean and variance of the outcome of investments, regardless of the distribution of the outcome of investments [8]. One measure that takes mean of return, and standard deviation of return into account is the Sharpe ratio [8]. The Sharpe ratio is defined by the following formula [8]:

$$\frac{\mu_r - r_f}{\sigma_r} \tag{26.9}$$

r_f is the risk free rate of interest (this is negligible in high frequency), and μ_r and σ_r are the mean and standard deviation of return respectively. A high Sharpe ratio implies a high mean return per unit risk and vice-versa.

26.3 Framework

Our framework develops trading algorithms of the form:

if *entry* rule for long position is met **then**
  Go Long
else if *exit* rule for short position is met **then**
  Go short
end if
if Long and (*exit* rule for long position is met) **then**
  Exit long position
  evaluate payoff
else if *Stop* rule for long position **then**
  Exit long position
  evaluate payoff
end if
if Short and (*Exit* rule for short position is met) **then**
  Exit short position
  evaluate payoff
else if *stop rule* for short position is met **then**
  Exit short position
  evaluate
end if

Our framework coevolves the *entry*, *exit*, and *stoploss* rules for long and short positions respectively. Each set of rule is a species on its own. We denote the species of *entry*, *exit*, and *stop loss* rules for long positions as $E_L^k, k \in \{1, 2, 3, \ldots, N\}$, $C_L^k, k \in \{1, 2, 3, \ldots, N\}$, and $S_L^k, k \in \{1, 2, 3, \ldots, N\}$ respectively. $E_S^k, k \in \{1, 2, 3, \ldots, N\}$, $C_S^k, k \in \{1, 2, 3, \ldots, N\}$, and $S_S^k, k \in \{1, 2, 3, \ldots, N\}$ is the notation for *entry*,*exit*, and *stop loss* rules for short positions. The transition table for the algorithm given above is in Table 26.1.

GE is used to evolve species within each population. Sexual reproduction is inter-species and we employ an implicit speciation technique within each species [9]. Rules are spatially distributed on notional toroid and sexually reproduce with rules within their *deme* [9]. This is akin to individuals sharing information with individuals within their social circle. The *deme* of a rule k is a set of individuals within the immediate vicinity of rule k on the imaginary toroid. Collaborators are chosen from other species based on an elitist principle [10]. For instance, when assessing a solution from the set $E_L^k, k \in \{1, 2, 3, \ldots, N\}$, the best from $C_L^k, k \in \{1, 2, 3, \ldots, N\}$, $S_L^k, k \in \{1, 2, 3, \ldots, N\}$, $E_S^k, k \in \{1, 2, 3, \ldots, N\}$, $C_S^k, k \in \{1, 2, 3, \ldots, N\}$, and $S_S^k, k \in \{1, 2, 3, \ldots, N\}$ are chosen for collaboration and the fitness attained is

Table 26.1 Transition table for *entry*, *exit*, and *stop loss* rules

Current position	E_L^k	C_L^k	S_L^k	E_S^k	C_S^k	S_S^k	Action
Long	X	0	0	X	X	X	Do nothing
Long	X	0	1	X	X	X	Close long position
Long	X	1	0	X	X	X	Close long position
Long	X	1	1	X	X	X	Close long position
Short	X	X	X	X	0	0	Do nothing
Short	X	X	X	X	0	1	Close short position
Short	X	X	X	X	1	0	Close short position
Short	X	X	X	X	1	1	Close short position
Neutral	0	X	X	0	X	X	Do nothing
Neutral	0	X	X	1	X	X	Open short position
Neutral	1	X	X	0	X	X	Open long position
Neutral	1	X	X	1	X	X	Donothing

X stands for ignore

assigned to rule k. The platform for collaboration and fitness evaluation is a trading algorithm. Each species asserts evolutionary pressure on the other and rules that contribute to the profitability of the trading algorithm attain high fitness and survive to pass down their genetic material to their offspring. On the other hand, rules that do not contribute are awarded low fitness and are eventually replaced by solutions with higher fitness.

Selection occurs at the population level such that for each species a tournament is performed and if the fitness of a rule is less than the fitness of its offspring then it is replaced by its offspring. This can be expressed formally using the following equation:

$$x = \begin{cases} y & \text{If } f_y > f_x \\ x & \text{otherwise} \end{cases} \tag{26.10}$$

26.3.1 Objective Function

The following assumptions are implicit in the fitness evaluation:

1. Only one position can be traded at any instant.
2. Only one unit can be traded at any instant.
3. The is no market friction (zero transaction cost, zero slippage, zero market impact). Arguably, since only one unit is traded at any instant, the effect of market impact can be considered to be negligible.

26.3.1.1 Sharpe Ratio

The Sharpe ratio is computed using Eq. 26.10. The objective is then to maximise:

$$\max \frac{\mu_{r^k}}{\sigma_{r^k}} \tag{26.11}$$

μ_r^k, and σ_r^k are the mean and standard deviation of trading algorithm $k, k \in \{1, 2, 3, \ldots, P\}$.

26.3.1.2 Expected Utility

The objective function when using utility functions is the expected utility which is calculated as follows:

$$f = E(U(W)) = \frac{1}{N} \sum_{j=1}^{30} \sum_{i=1}^{N} U(W_i, \theta_j) \tag{26.12}$$

where $U(W_i, \theta_j)$ is the utility of wealth at interval $i, \imath \in \{1, 2, 3, \ldots, N\}$ given the vector of parameters for the utility function θ_j, and N is the number of trading intervals. The utility for each interval is calculated for a range of parameter values (see Sect. 26.2 for parameter settings). The objective in the utilitarian framework can be formally expressed as follows:

$$\max_{\theta_j} E(U(W_i, \theta_j)), i \in \{1, 2, 3, \ldots N\} \tag{26.13}$$

26.3.2 Parameter Settings

The population size P of each species is set to 100 and the coevolutionary process is allowed to happen for a maximum number of generations, MaxGen = 200. The coevolutionary process is terminated after MaxGen/2 generations, if there is no improvement in the fitness of the elitist (best solution) of the best solutions from each species. The *deme* size for each species is set to 11. The grammar used in mapping the *entry*, and *exit* rules of the trading algorithms is shown in Table 26.2. The grammar used in mapping the *stop loss* rules of the trading systems is shown in Table 26.3. In our notation, O(t−n:t−1) represents a set of open prices, C(t−n:t−1) represents a set of closing prices, H(t−n,−1) represents a set of highest prices, and L(t−n:t−1) represents a set of lowest prices between t−n and t−1. O(t−n) represents the open price at t−n, C(t−n) represents the closing price at t−n, H(t−n) represents the highest price at t−n, and L(t−n) represents the lowest price at t−n. Where $n \in \{10, 11, 12, \ldots, 99\}$ and $t \in \{1, 2, \ldots\}$. sma(•), and ema(•) stand for simple, and exponential moving average respectively.

Table 26.2 Grammar for mapping E_L^k, E_S^k, C_L^k, and C_S^k

ϕ	Rule		n
<expr> ::	<binop> (<expr> , <expr>)		
	<rule>		(2)
<rule> ::	<var> <op> <var>	<var> <op> <fun>	
	<fun> <op> <fun>		(3)
<binop> ::	and, or, xor		(3)
<var> ::	H(t-<window>)	L(t-<window>)	
	O(t-<window>)	C(t-<window>)	
<op> ::	$>, <, =, \leq, \geq,$		(4)
<window> ::	<integer> <integer>		(1)
<integer> ::	1, 2, 3, 4, 5, 6, 7, 8, 9		(9)
<fun> ::	sma(H(t-<window> :t−1))	ema(H(t-<window> :t−1))	
	max(H(t-<window> :t−1))	min(H(t-<window> :t−1))	
	sma(L(t-<window> :t−1))	ema(L(t-<window> :t−1))	
	max(L(t-<window> :t−1))	min(L(t-<window> :t−1))	
	sma(O(t-<window> :t−1))	ema(O(t-<window> :t−1))	
	max(O(t-<window> :t−1))	min(O(t-<window> :t−1))	
	sma(C(t-<window> :t−1))	ema(C(t-<window> :t−1))	
	max(C(t-<window> :t−1))	min(C(t-<window> :t−1))	(15)

ϕ is the set of non-terminals, and n is the n is the number of rules for mapping the non-terminal ϕ

Table 26.3 Grammar for mapping S_L^k, and S_S^k

ϕ	Rule		n
<expr> ::	<preop> (<expr> , <expr>)		
	<rule>		(2)
<rule> ::	<rule> <op> <rule>	<var> <op> <var>	
	<var> <op> <fun>	<fun> <op> <fun>	
	<fun>	<var>	(6)
<preop> ::	min, max		(2)
<var> ::	H(t-<window>)		
	L(t-<window>)		
	O(t-<window>)		
	C(t-<window>)		(4)
<fun> ::	sma(H(t-<window> :t-1))	ema(H(t-<window> :t-1))	
	max(H(t-<window> :t-1))	min(H(t-<window> :t-1))	
	sma(L(t-<window> :t-1))	ema(L(t-<window> :t-1))	
	max(L(t-<window> :t-1))	min(L(t-<window> :t-1))	
	sma(O(t-<window> :t-1))	ema(O(t-<window> :t-1))	
	max(O(t-<window> :t-1))	min(O(t-<window> :t-1))	
	sma(C(t-<window> :t-1))	ema(C(t-<window> :t-1))	
	max(C(t-<window> :t-1))	min(C(t-<window> :t-1))	(15)
<window> ::	<integer> <integer>		(1)
<integer> ::	1, 2, 3, 4, 5, 6, 7, 8, 9		(9)

ϕ is the set of non terminals, and n is the number of rules for mapping the non-terminal ϕ

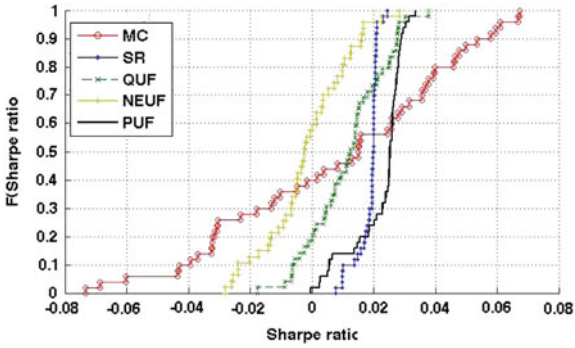


Fig. 26.1 Average out-of-sample Sharpe ratio of trading systems produced under the assumption of PUF, NEUF, QUF, and sharpe ratio fitness functions

The parameters of the power utility function (PUF), λ, and γ are sampled within the following ranges. $1<\lambda<2$, and $1<\gamma<35$ where λ controls the degree of loss aversion and γ controls the degree of risk aversion. The parameters of the negative exponential utility function (NEUF), a, b, and c were sampled from the following ranges $1<a<35$, $1<b<35$, and $1<c<35$. The parameter for the quadratic utility function (QUF) was sampled from the following ranges $1<b<35$.

26.4 Data

In this chapter, we use high frequency tick data for Amvesco for the period between 1 March 2007 and 1 April 2007 for our study. The data was compressed into a series of five minutely high, low, open, close prices proxy. The data was then divided into four blocks for k-fold cross validation [11].

26.5 Results and Discussion

In this section, we present the results obtained from producing trading algorithms under the assumption of power utility function (PUF), negative exponential utility function (NEUF), quadratic utility function (QUF), and Sharpe ratio (SR) as objective function. Utility is not comparable across different utility functions hence, analysis is performed directly on the returns obtained by the trading algorithms. We take the average of the performance of the trading algorithms across different blocks in accordance with k-fold cross validation. Furthermore, the trading algorithms developed are compared to the turtle trading systems (TTS) (see Sect. 26.2.1. The comparison will serve as a test for the hypothesis that trading systems developed using our framework are able to outperform the turtle trading system.

In addition, the trading algorithms developed are compared to a set of randomly initialised trading algorithms (MC). The randomly initialised trading algorithms

Table 26.4 Kruskal-Walliss ANOVA test results for out-of-sample average Sharpe ratios of agents produced under assumption of PUF, NEUF, QUF, and Sharpe ratio as fitness functions

χ^2	p-value
85.640	0.000

Table 26.5 Kruskal-Walliss ANOVA test results for the null hypothesis that the out-of-sample Sharpe ratios of agents produced under assumption of PUF, NEUF, QUF, and Sharpe ratio as fitness functions is the same as a set of random strategies (MC)

Objective function	PUF	NEUF	QUF	SR
χ^2	19.960	1.400	3.110	8.340
p-value	0.000	0.237	0.078	0.040

were mapped using the grammar used to coevolve our trading algorithms. The comparison will test if the performance of the trading systems can be reproduced by chance.

Figure 26.1 depicts the cumulative distribution function of the average Sharpe ratios of trading algorithms produced under the assumption of PUF, NEUF, QUF, and Sharpe ratio as fitness functions. Figure 26.1 suggests that, given the assumption of no budget constraints and frictionless markets, trading systems produced under the assumption of PUF, and SR have a better reward to risk ratio (Sharpe ratio) for the data set considered. A kruskal-Walliss ANOVA test for the null hypothesis that the Sharpe ratios of agents produced under assumption PUF, NEUF, QUF, and SR as fitness functions are the same was performed and the test results are shown in Table 26.4. The results in Table 26.4 show that, trading systems produced under the assumption of PUF, NEUF, QUF, and SR produce Sharpe ratios that are statistically different from each other. The results in Table 26.4 support the results in Fig. 26.1. Traditional investment theory postulates that provided investors have utility functions that satisfy the assumption of risk aversion and non-satiation then irrespective of their utility functions the mean and standard deviation of the outcomes of their investments are enough to summarise the outcomes of the distribution of outcomes. All the utility functions employed in this chapter satisfy the assumption of risk aversion and non-satiation. The results in Fig. 26.1; however, show that there is a difference between trading systems developed using different utility functions, and trading systems developed using the Sharpe ratio.

Table 26.5 shows results from the Kruskal-Wallis ANOVA test for the null hypothesis that, the Sharpe ratios of the agents produced under the assumption of PUF, NEUF, QUF, and SR as objective function are not different from a set of randoml strategies (MC). The results in Table 26.5 suggests that trading systems produced under the assumption of PUF, QUF, and SR produce Sharpe ratios that are significantly different from a set of randomly initialised strategies. This implies performance of these trading systems is highly unlikely to have resulted out of pure chance.

To test the hypothesis that, our framework can be used to produce trading algorithms that outperform the turtle trading system, the performance of the

Table 26.6 Sign test results for the null hypothesis that the out-of-sample Sharpe ratios of trading systems produced under assumption PUF, NEUF, QUF, and SR as fitness functions come from a continuous distribution with a median that is same as the Sharpe ratio of the TTS

Objective function	z-value	sign	p-value
PUF	1.839	18	0.000
NEUF	−6.418	1	0.000
QUF	−4.000	10	0.000
SR	−6.647	1	0.000

trading systems developed is compared to the performance of the turtle trading system using a sign test. Table 26.6 contains results from a sign-test for the null hypothesis that average out-of-sample Sharpe ratios of the trading systems developed under assumption of power utility function (PUF), negative exponential utility function (NEUF), quadratic utility function (QUF), and Sharpe ratio (SR) as objective function are not any different from the turtle trading system (TTS). The results in Table 26.6 suggests that trading systems produced under the assumption of PUF as objective function produced Sharpe ratios that are significantly better than the Sharpe ratio of the TTS.

26.6 Chapter Summary

Advancements in communication and computer technology has allowed trading systems to be programmed into computer programs that execute orders and this has gained a lot of popularity [2]. In this chapter, we introduced a method based on GE for coevolving technical trading rules for high frequency trading (see Sect. 26.3 for the method). Our results suggests our framework is capable of producing trading algorithms that outperform the turtle trading system under no budget constraint when using power utility function as objective function. The results in this chapter show that there is a significant difference between the performance of trading systems that were produced under the assumption of PUF, NEUF, QUF, and Sharpe ratio as objective function. This suggests that coevolutionary approach is highly sensitive to the objective function chosen.

References

1. Faith C (2003) The original turtle trading rules. http://www.originalturtle.org
2. Hendershott T, Jones CM, Menkveld AJ (2011) Does algorithmic trading improve liquidity. J Finance 66(1):1–33
3. O'Neill M, Brabazon A, Ryan C, Collins JJ (2001) Evolving market index trading tules using grammatical evolution. Appl Evolut Comput, Lect Notes Comput Sci 2037(2001):343–352

4. Anderson JA (2003) Taking a peek inside the turtle's shell. School of Economics and Finance, Queensland University of Technology, Australia
5. Adamu K, Phelps S (2010) Coevolution of technical trading rules for high frequency trading. Lecture notes in computer science and engineering, proceedings of the world congress on engineering, WCE 2010(1):96–101
6. Saks P, Maringer D (2009) Evolutionary money management. Lecture notes in computer science, Applications of evolutionary computing, vol 5484(2009). Springer, Heidelberg pp 162–171
7. Cuthbertson K, Nitzsche D (2004) Quantitative financial economics. 2nd edn. Wiley, Chichester pp 13–32 (chapter 1)
8. Amman H, Rusten B, (eds) (2005) Portfolio management with heuristic optimization. Advances in computational management sciences, vol 8. Springer, Berlin pp 1–37 (Chapter1)
9. Eiben AE, Smith JE (2003) Introduction to evolutionary computing. Springer, Berlin (Chapter 9)
10. Wiegand R, Paul C, Liles W, JongKenneth A De (2001) An empirical analysis of collaboration methods in cooperative coevolutionary algorithms. In: Proceedings of the genetic and evolutionary computation conference, Morgan Kaufmann Publishers
11. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Joint Conf Artif Intell 14(2):1137–1145

Chapter 27
High Performance Computing Applied to the False Nearest Neighbors Method: Box-Assisted and kd-Tree Approaches

Julio J. Águila, Ismael Marín, Enrique Arias, María del Mar Artigao and Juan J. Miralles

Abstract In different fields of science and engineering (medicine, economics, oceanography, biological systems, etc.) the false nearest neighbors (FNN) method has a special relevance. In some of these applications, it is important to provide the results in a reasonable time scale, thus the execution time of the FNN method has to be reduced. To achieve this goal, a multidisciplinary group formed by computer scientists and physicists are collaborative working on developing High Performance Computing implementations of one of the most popular algorithms that implement the FNN method: based on box-assisted algorithm and based on kd-tree data structure. In this paper, a comparative study of the distributed memory architecture implementations carried out in the framework of this collaboration is

J. J. Águila (✉) · E. Arias
Albacete Research Institute of Informatics, University of Castilla-La Mancha, Avda. España s/n, 02071 Albacete, Spain
e-mail: juliojose.aguila@alu.uclm.es

E. Arias
e-mail: enrique.arias@uclm.es

I. Marín · M. del Mar Artigao · J. J. Miralles
Applied Physics Department, University of Castilla-La Mancha, Avda. España s/n, 02071 Albacete, Spain
e-mail: ismael.marin@uclm.es

M. del Mar Artigao
e-mail: mariamar.artigao@uclm.es

J. J. Miralles
e-mail: juan.miralles@uclm.es

J. J. Águila
Depto. Ingeniería en Computación, Universidad de Magallanes, Avda. Bulnes, 01855 Punta Arenas, Chile

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_27,
© Springer Science+Business Media B.V. 2011

presented. As a result, two parallel implementations for box-assisted algorithm and one parallel implementation for the kd-tree structure are compared in terms of execution time, speed-up and efficiency. In terms of execution time, the approaches presented here are from 2 to 16 times faster than the sequential implementation, and the kd-tree approach is from 3 to 7 times faster than the box-assisted approaches.

27.1 Introduction

In nonlinear time series analysis the false nearest neighbors (FNN) method is crucial to the success of the subsequent analysis. Many fields of science and engineering use the results obtained with this method. But the complexity and size of the time series increase day to day and it is important to provide the results in a reasonable time scale. For example, in the case of electrocardiogram study (ECG), this method have to achieve real-time performance in order to take some prevention actions. With the development of the parallel computing, large amounts of processing power and memory capacity are available to solve the gap between size and time.

The FNN method was introduced by Kennel et al. [1]. Let $X = \{x(i) : 0 \leq i < n\}$ a time series. We can construct points (delay vectors) according to

$$y(i) = [x(i), x(i+\tau), \ldots, x(i+(d-1)\tau)] \tag{27.1}$$

where τ is the embedding delay and d is the embedding dimension [2]. The Takens embedding theorem [3] states that for a large enough embedding dimension $d \geq m_0$, the delay vectors yield a phase space that has exactly the same properties as the one formed by the original variables of the system. The FNN method is a tool for determining the minimal embedding dimension m_0. Working in any dimension larger than the minimum leads to excessive computation when investigating any subsequent question (Lyapunov exponents, prediction, etc.).

The method identifies the nearest neighbor $y(j)$ for each point $y(i)$. According to Eq. 27.2, if the normalized distance is larger than a given threshold R_{tr}, then the point $y(i)$ is marked as having a false nearest neighbor.

$$\frac{|x(i+d\tau) - x(j+d\tau)|}{||y(i) - y(j)||} > R_{tr} \tag{27.2}$$

Equation 27.2 has to be calculated for the whole time series and for several dimensions $d = \{1, 2, \ldots, m\}$ until the fraction of points, which must be lower than R_{tr}, is zero, or at least sufficiently small (in practice, lower than 1%).

While greater is the value of n (length of the time series), the task to find the nearest neighbor for each point is more computationally expensive. A review of

methods to find nearest neighbors, which are particularly useful for the study of time series data, can be found in [4]. We focused in two approaches: based on the box-assisted algorithm, optimized in the context of time series analysis by [5]; and the based in a kd-tree data structure [6, 7] developed in the context of computational geometry.

According to Schreiber, for time series that have a low dimension of embedding (e.g. up to the 10's), the box-assisted algorithm is particularly efficient. This algorithm can offer a lower complexity of $O(n)$ under certain conditions. By the other hand, accordingly with the literature if the dimension of embedding is moderate an effective method for nearest neighbors searching consists in using a kd-tree data structure [6, 7]. From the computational theory point of view, the kd-tree-based algorithm has the advantage of providing an asymptotic number of operations proportional to $O(n \log n)$ for a set of n points, which is the best possible performance for arbitrary distribution of elements.

We have applied the paradigm of parallel computing to implement three approaches directed towards distributed memory architectures, in order to make a comparative study between the method based on the box-assisted algorithm and the method based on the kd-tree data structure. The results are presented in terms of performance metrics for parallel systems, that is, execution time, speed-up and efficiency. Two case studies have been considered to carried out this comparative study. A theoretical case study which consists on a Lorenz model, and a real case study which consists on a time series belonging to electrocardiography.

The paper is organized as follows. After this introduction, a description of the considered approaches is introduced in Sect. 27.2. In Sect. 27.3, the experimental results are presented. Finally, in Sect. 27.4 some conclusions and future work are outlined.

27.2 Parallel Approaches

We selected two programs to start this work: the `false_nearest` program based on the box-assisted algorithm [8, 9]; and the `fnn` program based on a kd-tree data structure [10].

We employ the paradigm *Single-Program*, *Multiple Data* (SPMD by [11]) to design the three parallel approaches. A coarse-grained decomposition [12] has been considered, i.e. we have a small number of tasks in parallel with a large amount of computations. The approaches are directed towards distributed memory architectures using the Message Passing Interface [13] standard for communication purpose. Two approaches are based on the box-assisted algorithm and the another approach is based on the kd-tree data structure.

27.2.1 Approaches Based on Box-Assisted Algorithm

The box-assisted algorithm [5] considers a set of n points $y(i)$ in k dimensions. The idea of the method is as follow. Divide the phase space into a grid of boxes of side length ϵ. Each point $y(i)$ lies into one of these boxes. The nearest neighbors there are located in the same box or in one of the adjacent boxes. The `false_nearest` program is a sequential implementation of the FNN method based on this algorithm.

By profiling the `false_nearest` program in order to carry out the parallel approaches, four tasks were identified. Let X a time series, Y a set of points constructed according to Eq. 27.1, `BOX` an array that implements the grid of boxes (or mesh), and p the number of processes. Two parallel implementations were formed based on these four tasks:

Domain decomposition Time series X is distributed to the processes. Two ways of distribution have been developed: Time Series (TS) and Mesh (M). In a TS data distribution the time series is split into p uniform parts of length $\frac{n}{p}$, being n the length of the time series. In a M data distribution, each process computes the points that lie in its range of rows. The range of the mesh rows is assigned by $\frac{s}{p}$, where s is the size of the BOX.

Grid construction The `BOX` array is filled. Two ways of grid construction have been developed: S (Sequential) and P (Parallel). In a S construction each process fills the BOX sequentially, thus each one has a copy. In a P construction each process fills a part of the group of boxes located over a set of assigned mesh rows.

Nearest neighbors search Each process solves their subproblems given the domain decomposition way. In a TS data distribution each process uses the same group of points Y. In a M data distribution each process can use different groups of points.

Communication of results Processes use MPI to synchronize the grid construction and to communicate the partial results at the end of each dimension.

The approaches were called following the next nomenclature: `DM-P-M` meaning a **D**istributed **M**emory implementation considering that the grid construction is in **P**arallel and the time series is distributed according to the **M**esh; `DM-S-TS` meaning a **D**istributed **M**emory implementation considering that the grid construction is **S**equential and the **T**ime **S**eries is uniformly distributed to the processes.

We have introduced MPI functions into the source codes to obtain the programs that can be run into a distributed memory platform. The most important MPI functions used in these programs are as follows:

- `MPI_Reduce` Combines values provided from a group of MPI processes and returns the combined value in the MASTER process.
- `MPI_Allreduce` Same as `MPI_Reduce` except that the result appears in all the MPI processes.

Let p the total number of MPI processes. Each process has an identifier $q = \{0, 1, \ldots, p-1\}$. The process $q = 0$ is treated as MASTER and processes with $q \neq 0$ are treated as slaves. The next algorithm depicts the algorithmic notation for the `DM-P-M` approach:

Program *DM-P-M*(m, τ, X)
Input:
m = maximal embedding dimension to compute
τ = embedding delay
X = time series data record of length n
Output:
The fraction of false nearest neighbors for each dimension.
in parallel do:
begin
Computing bounds $first$ and $last$ to get same number of rows: s/p; `/* s is the mesh size.*/`
for $d = 1$ **to** m **do**
Setting the initial value of ϵ;
Setting the control variable $alldone$ to $FALSE$;
while $alldone = FALSE$ **and** $\epsilon < threshold$ **do**
Building $BOX(first : last)$ using ϵ, τ, d and X;
for each $y(i)$ **into** $BOX(first : last)$ **do**
Searching the nearest neighbor of $y(i)$;
if nearest(i) is found **then**
Computing if nearest(i) is a false nearest;
endif
endfor
synchronization Calling to MPI_Allreduce($alldone$);
`/* If` $alldone = TRUE$ `meaning that all nearest were founded.*/`
Updating ϵ;
endwhile
synchronization Calling to MPI_Reduce(fnn);
if $q = MASTER$ **then** Printing fnn;
endfor
end

The next algorithm depicts the algorithmic notation for the `DM-S-TS` approach:

Program $DM\text{-}S\text{-}TS(m, \tau, X)$
Input:
m = maximal embedding dimension to compute
τ = embedding delay
X = time series data record of length n
Output:
The fraction of false nearest neighbors for each dimension.
in parallel do:
begin
Computing bounds ini and end to get same number of data: n/p;
for $d = 1$ **to** m **do**
Setting the initial value of ϵ;
Setting the control variable $alldone$ to $FALSE$;
while $alldone = FALSE$ **and** $\epsilon < threshold$ **do**
Building BOX using ϵ, τ, d and X;
for each i **into** $y(ini : end)$ **do**
Searching the nearest neighbor of $y(i)$;
if nearest(i) is found **then**
Computing if nearest(i) is a false nearest;
endif
endfor
synchronization Calling to MPI_Allreduce($alldone$);
`/* If` $alldone = TRUE$ `meaning that all nearest were founded.*/`
Updating ϵ;
endwhile
synchronization Calling to MPI_Reduce(fnn);
if $q = MASTER$ **then** Printing fnn;
endfor
end

27.2.2 Approach Based on the kd-Tree Data Structure

A kd-tree data structure [6, 7] considers a set of n points $y(i)$ in k dimensions. This tree is a k-dimensional binary search tree that represents a set of points in a k-dimensional space. The variant described in Friedman et al, distinguishes between two kinds of nodes: internal nodes partition the space by a cut plane defined by a value of the k dimensions (the one containing a maximum spread), and external nodes (or buckets) store the points in the resulting hyperrectangles of the partition. The root of the tree represents the entire k-dimensional space. The `fnn` program is a sequential implementation of the FNN method based on this structure.

`fnn` program has been also analyzed by means of a profile tool before making the parallel implementation, identifying five main tasks. Thus, let X a time series, n the length of the time series, Y a set of points constructed according to Eq. 27.1, `KDTREE` a data structure that implements the kd-tree, p the number of processes, and $q = \{0, 1, \ldots, p-1\}$ a process identifier. For convenience we assume that p is a power of two. The parallel implementation called `KD-TREE-P` was formed based on these five tasks:

Global kd-tree building The first $\log p$ levels of `KDTREE` are built. All processors perform the same task, thus each one has a copy of the global tree. The restriction $n \geq p^2$ is imposed to ensure that the first $\log p$ levels of the tree correspond to nonterminal nodes instead of buckets.

Local kd-tree building The local `KDTREE` is built. In the level $\log p$ of the global tree are p nonterminal nodes. Each processor q builds a local kd-tree using the $(q+1)$th-node like root. The first $\log p$ levels are destroyed and `KDTREE` is pointed to local tree.

Domain decomposition Time series X is distributed to the processes. The building strategy imposes a distribution over the time series. Thus, the time series is split according to the kd-tree algorithm and the expected value of items contained in each local tree is approximately $\frac{n}{p}$.

Nearest neighbors search Each process solves their subproblems. Each process searches the nearest neighbors for all points in Y that are in the local `KDTREE`.

Communication of results Processes use MPI to communicate theirs partial results at the end of whole dimensions. The master process collects all partial results and reduces them.

The next algorithm depicts the algorithmic notation for the `KD-TREE-P` approach:

Program *KD-TREE-P*(m, τ, X)
Input:
m = maximal embedding dimension to compute
τ = embedding delay
X = time series data record of length n
Output:
The fraction of false nearest neighbors for each dimension.
in parallel do:
begin
`/* Each process has an identifier` q`.`
`The process` $q = 0$ `is treated as MASTER.`
`Processes with` $q \neq 0$ `are treated as slaves.`
`The total number of processes is` p`.*/`
for $d = 1$ **to** m **do**
Building delay vectors Y using X and τ;
Building first log(p) levels of $KDTREE$;

```
    Building (q + 1)th-node of KDTREE on level log(p);
     /* The (q + 1)th-node is the new root for
     KDTREE.*/
     for each y(i) into KDTREE do
       Searching the nearest neighbor of y(i);
       Computing if nearest(i) is a false nearest;
     endfor
     synchronization Calling to MPI_Reduce(fnn);
     if q = MASTER then Printing fnn;
    Free memory;
  endfor
end
```

27.3 Experimental Results

In order to test the performance of the parallel implementations, we have considered two case studies: the Lorenz time series generated by the equations system described in [14]; the electrocardiogram (ECG) signal generated by a dynamical model introduced in [15]. The Lorenz system is a benchmark problem in nonlinear time series analysis and the ECG model is used for biomedical science and engineering [16].

The parallel implementations have been run in a supercomputer called GALGO, which belongs to the Albacete Research Institute of Informatics [17]. The parallel platform consists in a cluster of 64 machines. Each machine has two processors Intel Xeon E5450 3.0 GHz and 32 GB of RAM memory. Each processor has 4 cores with 6,144 KB of cache memory. The machines are running RedHat Enterprise version 5 and using an Infiniband interconnection network. The cluster is presented as an unique resource which is accessed through a front-end node.

The results are presented in terms of performance metrics for parallel systems described in [12]: execution time T_p, speed-up S and efficiency E. These metrics are defined as follows:

- *Execution time* The serial runtime of a program is the time elapsed between the beginning and the end of its execution on a sequential computer. The parallel runtime is the time that elapses from the moment that a parallel computation starts to the moment that the last processing element finishes its execution. We denote the serial runtime by T_s and the parallel runtime by T_p.
- *Speed-up* is a measure that captures the relative benefit of solving a problem in parallel. It is defined as the ratio of the time taken to solve a problem in a single processing to the time required to solve the same problem on a parallel computer with p identical processing elements. We denote speed-up by the symbol S.
- *Efficiency* is a measure of the fraction of time for which a processing element is usefully employed; it is defined as the ratio of speed-up to the number of processing elements. We denote efficiency by the symbol E. Mathematically, it is given by $E = \frac{S}{p}$.

Table 27.1 Size of `BOX` for each value of p using a Lorenz time series

p	DM-P-M	DM-S-TS
1	8,192	8,192
2	4,096	4,096
4	2,048	4,096
8	2,048	4,096
16	2,048	2,048
32	2,048	2,048

Table 27.2 Size of `BOX` for each value of p using a ECG time series

p	DM-P-M	DM-S-TS
1	4,096	4,096
2	4,096	4,096
4	4,096	4,096
8	2,048	4,096
16	2,048	2,048
32	2,048	2,048

Let p the number of processors, the execution time of the approaches have been tested for $p = \{1, 2, 4, 8, 16, 32\}$, where $p = 1$ corresponds to the sequential version of the approaches. We used one million records of the time series to calculate the ten first embedding dimensions. We have obtained that the optimal time delay for Lorenz time series is $\tau = 7$ and for ECG signal is $\tau = 5$ using the mutual information method.

In order to obtain the best runtime of the approaches based in a box-assisted algorithm we found the best size of `BOX` for each value of p (Tables 27.1 and 27.2). The size of `BOX` defines the number of rows and columns for the grid of boxes. The values for $p = 1$ corresponds to the sequential version of the program `false_nearest`.

We have run ten tests to obtain the median value of the execution time T_p. In total 360 tests were performed. The performance metrics results are shown in Figs. 27.1 and 27.2.

Sequential kd-tree implementation shows a lower execution time than box-assisted approach, since the grid construction stage on box-assisted implementation in TISEAN is very expensive in terms of execution time.

The behavior of the Lorenz case study and the ECG case study is quite similar. Notice that, according to Figs. 27.1b and 27.2b, it is possible to appreciate a super-linear speed-up for kd-tree implementation when $p < 8$ and these performance decreases when $p > 8$. The super-linear speed-up is explained due to the fact that the cache memory is better exploited and that when the tree is split less searches have to be done at each subtree. With respect to the lost of performance, this situation is produced due to different causes. The first one is that, evidently, the overhead due to communications increases. Also, the most important cause is that the sequential part of our implementation becomes every time more relevant with respect to the parallel one.

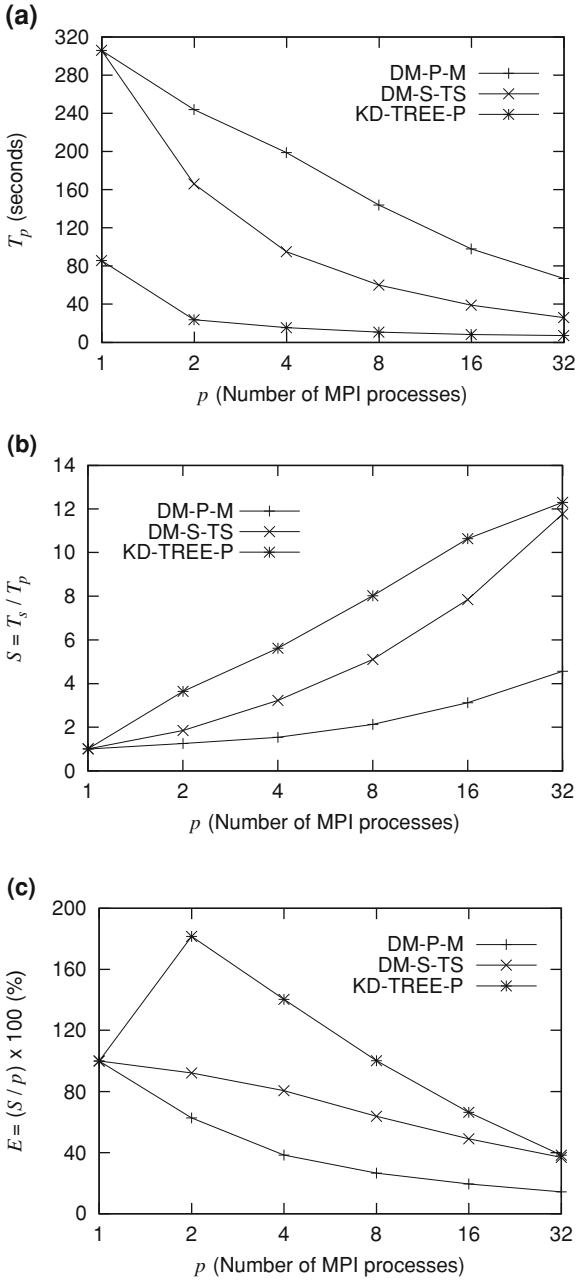


Fig. 27.1 Performance metrics for the Lorenz case study: **a** execution time; **b** speed-up; **c** efficiency

Considering only the box-assisted implementations, DM-S-TS is the box-assisted approach that provides the best results for the Lorenz attractor and the ECG signal. The reason is the very best data distribution with regard to DM-P-M.

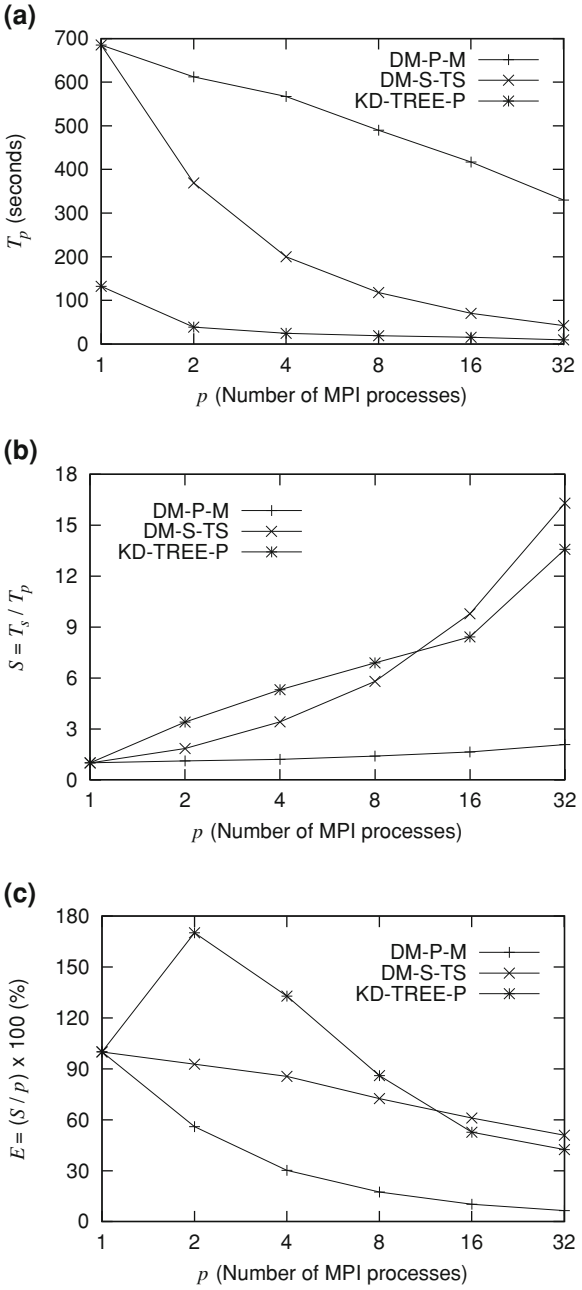


Fig. 27.2 Performance metrics for the ECG case study: **a** execution time; **b** speed-up; **c** efficiency

However, the reconstruction of the mesh is not parallelized in DM-S-TS implementation. So, the sequential part makes the reduction of execution time less significant when more CPUs are used. However, as the execution time of find

neighbors is increased (e.g. in larger times series data), this circumstance becomes very less important.

For Lorenz attractor, the DM-S-TS implementation is around 1.8 faster than the sequential program when it uses 2 CPUs, and around 12 when it uses 32 CPUs. This means that the efficiency for 2 CPUs is around 92% and decreases to 37% when using 32 CPUs. For ECG signal, the best box-assisted parallel implementation achieves a speed-up of around 16 when it is run on 32 CPUs of GALGO. Moreover, the time saving is around 93% using 2 CPUs and 51% using 32 CPUs. Unlike previous case, the efficiency of best implementation decreases more slowly.

An optimization of TISEAN has been used. It allows the best mesh size to be tuned for each case. In case of use original TISEAN (fixed mesh size), the reduction of execution time would be more important.

According to the experimental results, kd-tree-based parallel implementation obtains the best performance than the box-assisted-based parallel implementation, almost in terms of execution time, for both case studies. Due to the spectacular execution time reduction provided by the kd-tree-based parallel implementation, the performance in terms of speed-up and efficiency seems to be worst, with respect to the other approaches.

27.4 Conclusions

In this paper, a comparative study between the distributed memory implementations of two different ways to compute the FNN method have been presented, that is, the based on the box-assisted algorithm and the based on kd-tree data structure. To make this comparative study three different implementations have been developed: two implementations based on box-assisted algorithm, and one implementation based on kd-tree data structure.

The most important metric to consider is how well the resulting implementations accelerate the compute of the minimal embedding dimension, which is the ultimate goal of the FNN method. In terms of the execution time, the parallel approaches are from 2 to 16 times faster than the sequential implementation, and the kd-tree approach is from 3 to 7 times faster than the box-assisted algorithm.

With respect to the experimental results, the kd-tree-based parallel implementation provides the best performance in terms of execution time, reducing dramatically the execution time. As a consequence, the speed-up an efficiency are far from the ideal. However, it is necessary to deal with more case studies of special interest for the authors: wind speed, ozone, air temperature, etc.

About related works, in the context of parallel implementations to compute FNN method, the work carried out by the authors could be considered as the first one. The authors are working also on considering shared memory implementations using Pthreads [18, 19] or OpenMP [20, 21], and hybrid MPI+Pthreads or MPI+OpenMP parallel implementations. Also, as a future work, the authors are considering to develop GPU-based parallel implementation of the algorithms considered in this paper.

To sum-up, we hope that our program will be useful in applications of nonlinear techniques to analyze real time series as well as artificial time series. This work represents the first step of nonlinear time series analysis, that it is becomes meaningful when considering ulterior stages on the analysis as prediction, and when for some applications the time represents a crucial factor.

Acknowledgments This work has been supported by National Projects CGL2007-66440-C04-03 and CGL2008-05688-C02-01/CLI. A short version was presented in [22]. In this version, we have introduced the algorithmic notation by the parallel implementations.

References

1. Kennel MB, Brown R, Abarbanel HDI (1992) Determining embedding dimension for phase space reconstruction using the method of false nearest neighbors. Phys Rev A 45(6): 3403–3411
2. Fraser AM, Swinney HL (1986) Independent coordinates for strange attractors from mutual information. Phys Rev A 33(2):1134–1140
3. Takens F (1981) Detecting strange attractors in turbulence. In: Rand DA, Young L-S (eds) Dynamical systems and turbulence, Warwick 1980. Springer, New York, pp 366–381
4. Schreiber T (1995) Efficient neighbor searching in nonlinear time series analysis. Int J Bifurcation Chaos 5:349
5. Grassberger P (1990) An optimized box-assisted algorithm for fractal dimensions. Phys Lett A 148(1–2):63–68
6. Bentley JL (1975) Multidimensional binary search trees used for associative searching. Commun ACM 18(9):509–517
7. Friedman JH, Bentley JL, Finkel RA (1977) An algorithm for finding best matches in logarithmic expected time. ACM Trans Math Software (TOMS) 3(3):209–226
8. Hegger R, Kantz H, Schreiber T (1999) Practical implementation of nonlinear time series methods: the TISEAN package. Chaos 9(2):413–435
9. Hegger R, Kantz H, Schreiber T (2007) Tisean: nonlinear time series analysis. http://www.mpipks-dresden.mpg.de/∼tisean
10. Kennel MB (1993) Download page of `fnn` program ftp://lyapunov.ucsd.edu/pub/nonlinear/fns.tgz
11. Darema F (2001) The spmd model: past, present and future. In: Lecture notes in computer science, pp 1–1
12. Grama A, Gupta A, Karypis G, Kumar V (2003) Introduction to parallel computing. Addison-Wesley, New York
13. Message Passing Interface http://www.mcs.anl.gov/research/projects/mpi
14. Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20(2):130–141
15. McSharry PE, Clifford GD, Tarassenko L, Smith LA (2003) A dynamical model for generating synthetic electrocardiogram signals. IEEE Trans Biomedical Eng 50(3):289–294
16. ECGSYN (2003) Ecgsyn: a realistic ecg waveform generator. http://www.physionet.org/physiotools/ecgsyn
17. Albacete Research Institute of Informatics, http://www.i3a.uclm.es
18. Mueller F (1999) Pthreads library interface. Institut fur Informatik
19. Wagner T, Towsley D (1995) Getting started with POSIX threads. Department of Computer Science, University of Massachusetts
20. Dagum L (1997) Open MP: a proposed industry standard API for shared memory programming. OpenMP.org

21. Dagum L, Menon R (1998) Open MP: an industry-standard API for shared-memory programming. IEEE Comput Sci Eng 5:46–55
22. Águila JJ, Marín I, Arias E, Artigao MM, Miralles JJ (2010) Distributed memory implementation of the false nearest neighbors method: kd-tree approach versus box-assisted approach. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, WCE 2010, 30 June–2 July, London, UK, pp 493–498

Chapter 28
Ethernet Based Implementation of a Periodic Real Time Distributed System

Sahraoui Zakaria, Labed Abdennour and Serir Aomar

Abstract This work presents the realization of a platform for testing and validating distributed real time systems (DRTS), by following a methodology of development. Our main contribution remains the realization of an industrial communication bus (*FIP*: *Factory Instrumentation Protocol*) implemented on an Ethernet platform. It focuses on improving the response time of the bus. For that, we use a deterministic implementation of *FIP's* services (variables identification, transmission functions, …) by exploiting the TCP/IP stack. The periodic communications are monitored by real time periodic threads, run on RTAI kernel.

28.1 Introduction

Real time and distributed industrial systems development often rests on an appropriate methodology. Their implementation may be based on using a programming language or on using a fast prototyping tool that involves simulators, code generators and hardware in the loop. So, the design follows a model, architecture and a language appropriate to the applied methodology. The validation step of such systems in particular, requires platforms and middlewares to distribute the controls, the computation or the data. These platforms insure the management of the field buses.

S. Zakaria (✉) · L. Abdennour · S. Aomar
Computer Science Department, EMP, BP 17 BEB, Algiers, Algeria
e-mail: z.sahraoui@gmail.com

L. Abdennour
e-mail: a-labed@hotmail.com

S. Aomar
e-mail: aoser10@gmail.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_28,
© Springer Science+Business Media B.V. 2011

Distributed real time applications must satisfy two conditions to communicate data: determinism and reliability. Conventionally, industrial local area networks or any networks in hostile environment (engine of a vehicle) use field buses as the controller area network (*CAN*) and the *FIP* buses to fit these two requirements. The *FIP* field bus is a platform offering a configuring interface allowing a station[1] to take place in the *FIP* network.

Hence, it can produce and consume periodic or aperiodic variables, send and receive messages with or without connection or assure the arbitration function of the bus. The arbitrator holds the list of variables which are created on configuration. For each variable the producer and its periodicity of use on the bus are defined [1, 2].

Industrial networks tend to exploit the possibilities of the ETHERNET, which has been proven to be well adapted since it has a lower performance/cost ratio.

In the present work we describe a platform that we have developed, to test the DRTS. We have adopted the methodology proposed by *Benaissa* and *Serir* in [3]. It is an approach of development based on a descendent hierarchical functional decomposition of a control system. The primitive processes of the system are specified by elementary grafcets that communicate and synchronize using messages. The Grafcets processes may refer to components of the same site or of different sites (distributed). The link between sites is provided by a communication system similar to the *FIP* field bus with a deterministic, periodic and synchronous behavior. Our contribution concerns a *FIP* configuration on *ETHERNET*.

So, we have projected the *FIP* bus architecture (protocols and services) on its counterpart, the Ethernet communication system together with *TCP/IP*. This aims at checking the specification of the resulting *FIP* bus to fit all the mechanisms of *FIP*: the use of broadcasting in the communication system, acknowledgment of variables identifiers by the different sites and errors verification services (physical, link and application layers of *OSI*). Consequently, the validation will rest on performance of the real time micro kernel *RTAI*[2] we have used.

We will analyze the *TCP/IP/Ethernet* model in Sect. 28.2. In Sect. 28.3 we describe the approach of design of the proposed *FIP* field bus. In Sect. 28.4, we implement an example of a distributed real time application specified according to the methodology. Section 28.5 resumes and discusses the results of the tests. We conclude in Sect. 28.6.

28.2 Related Works

Data exchange between and within field buses requires high rate communications. However, the majority of the field buses such as *WorldFip*, *FF*, *Profibus*, *P-Net* and *CAN* suggest insufficient rates (31.25–5 Mbps). To overcome this problem,

[1] Part of the bus that can be, a computer, an automaton, a sensor or an actuator.

[2] Open source, preemptible without latency problems of operating system calls and has a known gigue.

high rate networks as *FDDI* and *ATM* were proposed and their capability to support strict time constraints and soft real time has been evaluated. However, these solutions didn't meet expectations because of their high costs and complexity of implementation [4].

Some improvements have been operated on the Ethernet to help it support communications constrained by the time. So, the Ethernet protocol is modified or a deterministic layer is implemented over the MAC sublayer. One solution consists in using the *TDMA* strategy. But it has the drawback to waste the time slots of idle stations (no transmissions). As examples, we can cite the *P-CSMA* (Prioritized *CSMA*), *RTNET* of *RTAI* micro kernel. The *PCSMA* (Predictable *CSMA*) technique is data scheduling oriented where all real time data are assumed to be periodic. Though it avoids time waste, it has an overhead in its off line scheduling. We can also find techniques based on the modification of the binary exponential backoff (*BEB*) algorithm, like *CSMA/DCR* (*CSMA* deterministic collision avoidance) which uses a binary tree research instead of the non deterministic *BEB* [5]. Indeed, such techniques may support strict as well as soft real time applications by changing the basic structure of the Ethernet. Moreover, adding a deterministic layer upon MAC, may lead to the same result. Among these solutions, we have the Virtual Time protocol (VTP), the Window Protocol (WP) and the traffic smoothing (TS) [4].

Middleware-based protocols of communication have been recently proposed for applications in automation. They are implemented either on *TCP*, like *Modbus TCP* and *ProfiNet* or on *UDP*, such as *NDDS* (*Network Data Delivery Service*) [6, 7]. We end the list of technical solutions by *Avionics Full DupleX* switched Ethernet base 100 TX (*AFDX*).

28.3 Proposed Approach (FIP over ETHERNET/IP)

The platform design is the result of *Ethernet/TCP/IP* model and its counterpart *FIP's* layers analysis.

28.3.1 Analysis of the Physical Layer

The *WorldFIP* norm defines three transmission rates: 31,25 kbit/s with a bit transmission time of $Tbit = 32$ µs, 1 Mbit/s with $Tbit = 1$ µs and 2,5 Mbit/s with $Tbit = 400$ ns [1]. So, the Ethernet rate varies between 10 Mbit/s with $Tbit = 512$ µs and 10 Gbit/s with $Tbit = 512$ ns [8]. The Manchester II coding is used for Ethernet with 10 Mbit/s and for the *WorldFIP*. For instance, Ethernet 100BaseTx uses the 4B/5B coding which limits the *Tbit* to 8 ms [8]. Note that the effects of noise on the *FIP* bus are similar to those on Ethernet categories which use short *Tbit* and low modulation speed.

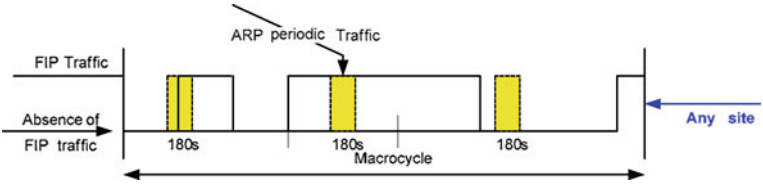


Fig. 28.1 Computation of the delay for the periodic traffic

PRE	DDT	Control	Data	CRC	DDT
8 bits	8 bits	1 octet	X octets	2 octets	8 bits

Fig. 28.2 Link layer frames' format of *WorldFIP* (*CENELEC EN61158-2 norm of FIP*) [1]

28.3.2 Analysis of the Data Link Layer

The first step consists in analyzing the protocols *ARP*[3] (address resolution protocol) and *LLC* (logical layer control) of *TCP/IP* and the services that may affect the traffic, in order to preserve a deterministic behavior. For our experiments, we used a LAN with one HUB and two PCs on which we installed a linux system provided with a traffic analyzer (*Ethereal*).

For the Ethernet traffic capture, we noticed that in absence of traffic, three queries may occur on the network: two *LLC* control queries (initialization), two *ARP* with the sender address (generated by default every 180 s) and active services queries. Furthermore, when we replaced the Hub by a switch, we noticed that it also uses the LLC protocol to initialize the network (*SSAP* query: Spanning tree *BPDU*[4] command with a forward delay of 15 ns).

To avoid non-determinism of the protocol, it suffices to fix the duration *Base-Reachable-Time* at an unreachable value. But, wasted time to ensure a deterministic emission or reception (periodic queries of identification, every 180 s) can be bounded by $(n + 1)$ *Tra*. In this relation, n is the ratio between duration of an emission and the period of an *ARP* query and *Tra* is the transmission time of an *ARP* query (Fig. 28.1).

28.3.3 Benefit of the Data Link Layer

The *FIP* frame format on the link level (Fig. 28.2) involves a control byte to code the frame's type (*ID_DAT, RP_DAT, ID_RQI, RP_ACK* …), data bytes (128 bytes

[3] *ARP*: protocol of layer three which makes the correspondence between Internet logical addresses and MAC addresses.

[4] Used by switches and routers to avoid loops on a WAN.

Fig. 28.3 Ethernet II frame

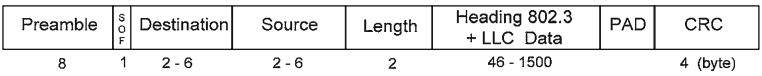


Fig. 28.4 The Ethernet IEEE 802.3 frame

or 256 octets) and two bytes allowing a receiver to check the integrity of the received frame [1].

Source and destination *MAC* addresses of Ethernet frames [5] (Figs. 28.3 and 28.4) have no role in the specification, since sites' identification in *FIP* bus has no interest at this level.

Using Ethernet *CSMA/CD* protocol, transmission errors are not detected through the absence of acknowledgment, but through interference. In *FIP* implementation over *ETHERNET*, the temporization is assured implicitly. However the frames (identifier, variables response, query response …) are processed at transport and application layers. Error control is required by the *CRC* for both *FIP* and Ethernet using the same code.

28.3.4 Benefit of the Network Layer (IP)

In the *OSI* standard, the *TCP/IP* protocol offers routing operations. So, interconnection between any pair of machines is possible. But in *FIP* network, sites identification is implicitly provided by the identifiers of variables to be transmitted. Recall that client sites of *FIP* system are synchronized only by variables identifiers and *IP* address will not give information on variables identifiers. Consequently, in our design, sites participating to the exchange of a variable are implicitly identified as producer of this variable via the broadcasting principle. The use of HUBs offers implicitly the broadcasting possibility. But, if we use switches, an appropriate configuration is necessary (configure ports on promiscuous mode).

28.3.5 Benefit of the UDP Layer

UDP protocol has been the unique solution for many tools of real time applications implementation. The nature of *UDP* datagrammes is ideal for sending fragments of data generated by such applications. It is selected for the speed of communication between its clients. It uses a simplified structure of the header, which restricts to the fields shown in Fig. 28.5. The checksum of the header is computed as for *IP* packets.

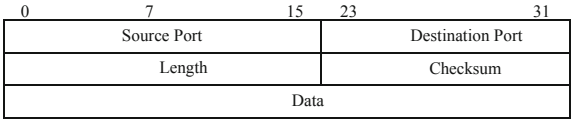


Fig. 28.5 The *UDP* datagramme or message format

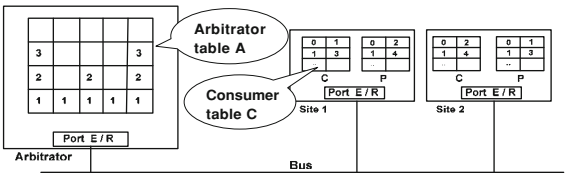


Fig. 28.6 Architecture of the adopted transport level

28.4 Adopted Architecture for the Transport Level

Architecture of Fig. 28.6 shows the solution we have chosen among much other architecture. Use of tables *P* and *C* to structure producer and consumer variables, simplifies managing variables at the processing step. Different port numbers are used for the arbitrator and the producer–consumer sites, to separate messages intended to identify variables and those which contain variables. Using a unique port instead of more than one at each producer/consumer site allows synchronous design of production and consumption functions.

A producer site receives identifiers *ID_DAT* of the variables to be produced. The consumer detects the arrival of these frames in order to enable an internal temporizer. If this temporizer expires, the station considers the next frame only if it has the emitting port of the arbitrator.

Recall that messages exchange on the *FIP* bus is done in point to point or in multipoint on the same segment. Two addresses of 24 bits (source and destination) allow coding the number of the segment of the application entity and its address on this segment. Hence, *IP* addressing may be used to perform these transactions.

28.5 Maximal Transferring Time (Critical Time)

Real time and distributed applications impose temporal constraints tasks achievement; these constraints will have direct impact on the exchanged messages between tasks located on different processors. In real time applications, tasks may have or not temporal constraints as well as the exchanged messages between them.

As indicated on Fig. 28.7, the transferring time of a message is composed of several intermediate times which are summarized in Table 28.1. If we note *Dt* as the duration of a transaction, then:

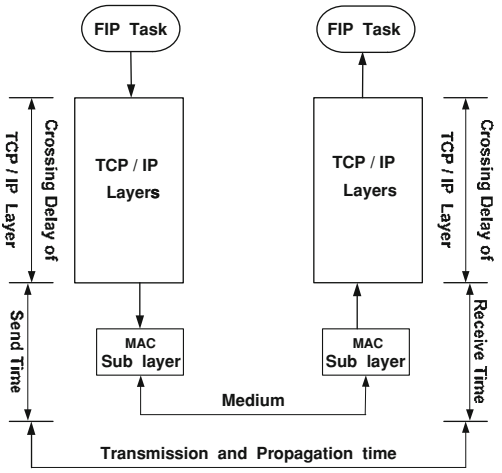


Fig. 28.7 Transmission times

Table 28.1 Notation used for transmission times

Identifiers transmission time	Notation	Variables transmission time	Notation
Identifier sending (*FIP* arbitrator task)	*Tta*	Variable sending time (*FIP* production task)	*Ttp*
Latest *Sendto*(*socket,UDP*…)	*Tst*	Latest *Sendto*(*socket,UDP*…)	*Tst*
MAC emission	*Tsm*	MAC emission time	*Tsm*
Transmission on the medium	*Tp*	Transmission on the medium	*Tp*
MAC reception	*Trm*	MAC reception time	*Trm*
Recvfrm(*Socket, UDP*…) *function*	*Trt*	*Recvfrm*(*Socket, UDP*…) *function*	*Trt*
Acknowledgement (*FIP* producer task)	*Trp*	Checking (*FIP* arbitrator task)	*Trp*

$$Dt = Tta + 2\,Tst + 2\,Tsm + 2\,Tp + 2\,Trm + 2\,Trt + 2\,Trp + Ttp\ldots \quad (28.1)$$

Given the fact that in our protocol, service is carried out by sources of indeterminism, which are all periodic, the maximal time of periodic transaction is computed as:

$$Dt \leq Tta + 2Tst + 2Tsm + 2Tp + 2Trm + 2Trt + 2\,Trp + Ttp + (n+1)\,Tra\ldots \quad (28.2)$$

28.6 Implementation of the Communication System

28.6.1 Hardware and Software of the Platform

Each component of our case materialized by a PC provided with an Ethernet network interface controller 100BaseT, is considered as a site. On each site the *RTAI* system is installed. We have used four: three are *Pentium 4* with CPU of

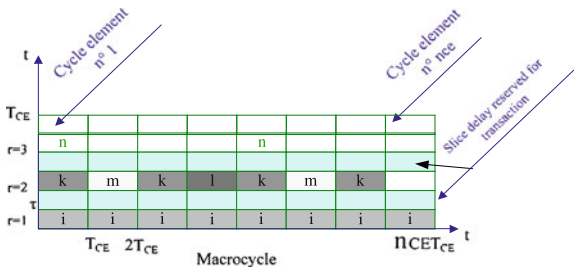


Fig. 28.8 Modified macrocycle

2.39 GHz; the fourth is a *Pentium 2* with CPU of 233 MHz. The first machine plays the role of arbitrator, the second and the fourth the role of producers–consumers (*site 1* and *site 2*). The third has a monitoring role by analyzing traffic using the *Ethereal* tool.

28.6.2 Our Arbitration Function

Variables' identifiers are scheduled on a macrocycle as follows:

- associate a task to each identifier;
- each task must be triggered periodically in the macrocycle at a precise time of the elementary cycle to broadcast the identifier;
- arbitrator executes tasks thanks to a preemptive scheduler with priority. A task is elected by scheduler only if all its preceding tasks have been achieved;
- remaining time after execution of all the tasks in a microcycle will be used for aperiodic exchange;
- awakening date of a periodic task *i* is computed by taking into account the total transfer time of all the previous transactions of a microcycle (Fig. 28.8).

28.6.3 The Producer–Consumer Function

We will specify production and consumption tasks of a site. The first task to be executed is the production function, because, the client site have first of all to wait for an eventual identifier of a variable using the primitive (*recvfrom*). Then, the producer task scrutinizes its table to check if it is concerned by the variable associated to this identifier in which case it broadcasts the variable. If the site is not the producer, the same task scrutinizes its consumption table to check if it has to receive the variable on the same port using the same primitive. On the other hand, in the consumption processing, the task enables an internal temporizer to confirm the frame loss at expiration of this temporizer and assure the global order of the system.

28.6.4 Use of Real Time FIFO Mechanism

Technically, it was not possible to compile a new Ethernet network driver over *RTAI*. So, *ETHERNET* of Linux system is used via the mechanism of real time queues (*rt_FIFO*), to communicate between ordinary processes and *RTAI* real time processes. The arbitrator creates two *rt_FIFOs* for its services; the reason is that, the primitive (*rtf_get*) used to read variables will use another mechanism of asynchronous nature. This primitive has been put in a function that we have called monitoring.

28.6.5 Monitoring Function

This function focuses on the variables exchange via real time queues and computes the transaction time. It is automatically enabled by the arrival of a variable in the queue (linux process has inserted the variable in the queue). This mechanism is assured by *rtf_create_handler*(*fifo, monitoring_func*) primitive. Hence, unlike the *FIP* bus variables, the *rt_FIFOs'* buffer has no refreshment and promptitude problem.

28.6.6 Schedulability

In our implementation we used an arbitrator which involves a set of *RTAI* periodic tasks and another function for the producer–consumer site. The latter is sequentially executed and respects the order of a *FIP* transaction. The fact that arbitrator tasks are periodic makes it possible to apply the schedulability criterion of formula (28.3) and to compute maximal times of transactions execution.

$$\sum_{i=1}^{i=n} DT_i/PT_i \leq 1\ldots \tag{28.3}$$

DT_i: ith FIP periodic transaction time,
PT_i: ith FIP period transaction.

28.7 A Case Study

In this example, the application is composed of two distributed grafcets in the communication system (Fig. 28.9).

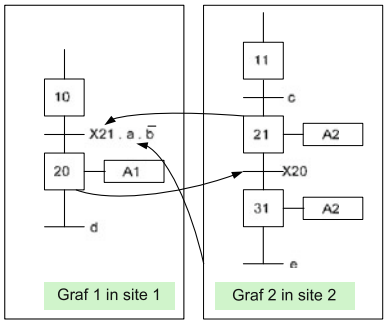


Fig. 28.9 Example of distributed grafcets

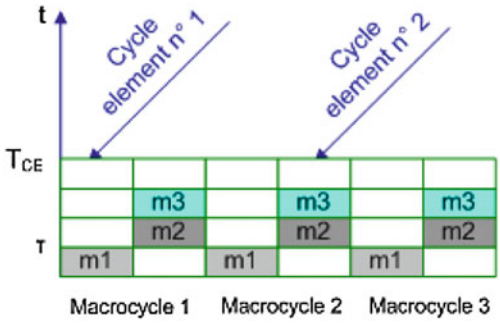


Fig. 28.10 Arbitrator table

Communication between the two grafcets requires transmission of input *a*(*m1*) and state *X21*(*m2*) of *site 2* to *site 1* and transmission of state *X20*(*m3*) from *site 1* to *site 2*, for each period of the macrocycle. Message *m1* is transmitted during the first elementary cycle. Messages *m2* and *m3* are transmitted during the second one in the order *m2*, *m3*.

To assure a coherent arbitration, we have bounded the time of a transaction, and consequently, the time to be added to the periods of variables. So, we compute the values resulting from subtractions between the clock value read after every sending and the corresponding value of the clock, sent by the monitoring function. Then, the upper bound is the maximum of the obtained results (Fig. 28.10).

28.7.1 Experimental Results

To compare our solution to the original *FIP*, we have considered the parameters: time of transactions and the duration of production function.

Since we have initialized the period, in timer tick, (periodic mode) to 119 ticks or 100,000 ns, a time value in tick of the clock is converted to nanoseconds by multiplying it by 840.336.

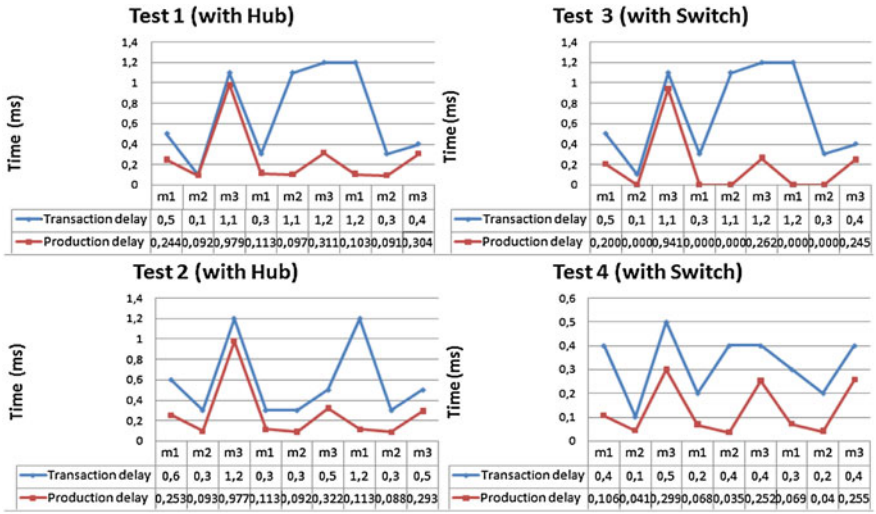


Fig. 28.11 Interpretation result

Table 28.2 Transaction times in milliseconds

	Test 1	Test 2	Test 3	Test 4
Average	0.460	0.600	0.700	0.322
Maximum	1.2	1.2	1.2	0.5
Minimum	0.3	0.3	0.1	0.1

We give bellow two sets of results (Fig. 28.11), corresponding respectively to a setup with a 10BaseT HUB and with a 100BaseT switch. Notice that we have used cables of *UTP5* category.

We have estimated the time used by a producer to produce the frame response of variable m_i and the associated propagation time.

Values sent by the identifier sending task and those sent by the monitoring function are given in tick of clock. The results of subtractions between the durations are converted into seconds.

28.7.2 Discussions

Table 28.2, gives an idea about some measured times.

It is obvious that transaction times may be lowered using a switch.

The maximal values are almost equivalent for all the tests and vary between 0.5 and 1.2 ms. These results are due to the fact that the production times are often less than half of transaction times, which explains the slowness and inderminism of emission and reception function of linux arbitrator.

Table 28.3 Sample of the original FIP scan speed [3]

Scanned variables	Variable size (bytes)	Scanned variables	Variable size (bytes)
320	1	181	16
304	2	123	32
277	4	75	64
235	8	42	128

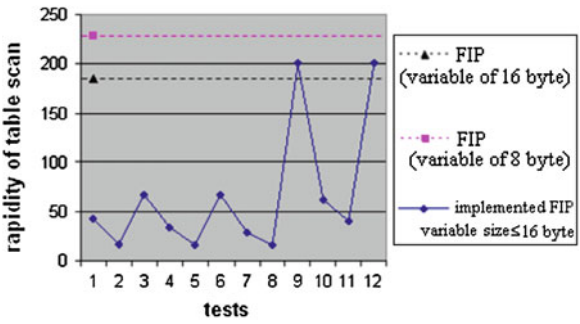


Fig. 28.12 Comparison of scrutation speeds

For example, if we take the value 1.2 ms which represents time of the seventh transaction of test 3, and the value 0.047 ms the time of its production. We notice that delay is due to emission and reception functions of the arbitrator (Fig. 28.11).

Another example, concerns the maximal value obtained in *test 1*, it corresponds to the transaction of variable *m3*. This value of 0,979 ms gives the production time of the variable by the slowest machine (*site 2*).

28.7.3 Speed of Variables Scrutation

Note that *FIP* Network at 2.5 Mbps, with a reversal time of 10 ms, in an element cycle of 20 ms, we can scrutinize (Table 28.3).

From Table 28.2, we can get the interval of variation for arbitrator scan speed (Fig. 28.12). We can notice that for the switch, the scan speed may reach 200 variables per 20 ms. However, for the *FIP* this speed is reached if the size of variables decreases from 16 to 8 bytes.

28.7.4 Useful Bit Rate

The useful bit rate is the ratio of the effective information and the duration of a transaction. For variables of size *L: 1 < L < 16* bytes, a *MAC* frame has always a size of 64 bytes. But, if the variables' size *L* is greater than 16 bytes, the size of the *MAC* frame will vary between 64 and 1500 bytes. Nevertheless, transmitted

Table 28.4 Comparison between the standard FIP and implemented FIP

	Useful data (byte)	Transmitted data (byte)	Transaction time (ms)	Useful bit rate (Kbps)	Efficiency (%)
FIP	4	19 or (16 + 4)	0.072	444.44	17.77
	8	23	0.084799	754.72	19.27
	16	31	0.013800	1159.42	32.32
	32	47	0.020200	1584.16	48.85
	64	79	0.033000	1939.39	65.64
	128	143	0.058600	2184.30	79.25
Implemented FIP	4	64	0.099999	5120	05.12
	8	64	0.099999	5120	05.12
	16	64	0.099999	5120	05.12
	32	79 or (47 + 32)	0.099999	6320	06.32

information is segmented into frames of 1500 bytes if the variables' size exceeds 1453 bytes.

28.7.5 Efficiency

We now compute the efficiency of our solution as the ratio between emission time of effective information and the duration of the transaction. It is equivalent to the useful bit rate and the transmission rate ratio.

Table 28.4 compares the *FIP's* [2] efficiency and that of our implementation. To complete the table, we deduce *FIP* transaction time from the ratio: length of useful information and useful bit rate.

To compute the useful bit rate, we take the global minimum of all the durations (tests of previous example). The tests on the implemented *FIP*, with a variable of 32 bytes gave the same minimums.

This table gives an idea about the margin that we have on the size of data that we can transmit in a transaction. Hence, efficiency is increased if we add other services (aperiodic variable exchange).

28.8 Conclusions and Future Work

We have compared the results obtained for the distributed grafcets and the standard *FIP* on a practical example. The comparison was concerned with the scanning speed of arbitration table and the computation of the communication system efficiency in its cyclic part. Results obtained using switched Ethernet (*switches*) show that arbitrator of the implemented *FIP* can scrutinize its arbitration table faster than the standard *FIP*.

The implemented platform confirms the goodness of the distributed grafcet model. It constitutes by itself a new design for implementation of distributed real time systems which can be qualified as distributed systems for field data base management.

References

1. WorldFIP tools FIPdesigner hlp technologies (2000) L M 2 - C N F - 2 - 0 0 1 – D, 12 Jul
2. WorldFip Protocole (1999) European standard, En 50170. http://www.WorldFIP.org
3. Benaissa M (2004) GRAFCET based formalism for design of distributed real time systems. Master Thesis, EMP Bordj-El-Bahri, Algiers, Algeria (in French)
4. Wang Z, Song YQ, Chen JM, Sun YX (2001) Real time characteristics of Ethernet and its improvement. In: Proceeding of the world congress on intelligent control and automation, June
5. Pujolle G (1997) Networks, 2nd edn. Eyrolles (in French)
6. Venkatramani C, Chiueh T, Supporting real-time traffic on Ethernet. In: 1052-8725/94 IEEE
7. Doléjs O, Hanzalék Z (2003) Simulation of Ethernet for real time applications. In: IEEE, ICIT—Maribor, Slovenia
8. Telecommunication and Networks, Claude Sevin, Dunod 2006 (in French)

Chapter 29
Preliminary Analysis of Flexible Pavement Performance Data Using Linear Mixed Effects Models

Hsiang-Wei Ker and Ying-Haur Lee

Abstract Multilevel data are very common in many fields. Because of its hierarchical data structure, multilevel data are often analyzed using Linear Mixed-Effects (LME) models. The exploratory analysis, statistical modeling, and the examination of model-fit of LME models are more complicated than those of standard multiple regressions. A systematic modeling approach using visual-graphical techniques and LME models was proposed and demonstrated using the original AASHO road test flexible pavement data. The proposed approach including exploring the growth patterns at both group and individual levels, identifying the important predictors and unusual subjects, choosing suitable statistical models, selecting a preliminary mean structure, selecting a random structure, selecting a residual covariance structure, model reduction, and the examination of the model fit was further discussed.

29.1 Introduction

Longitudinal data are used in the research on growth, development, and change. Such data consist of measurements on the same subjects repeatedly over time. To describe the pattern of individual growth, make predictions, and gain more insight

H.-W. Ker (✉)
Department of International Trade, Chihlee Institute of Technology, Taipei, 220, Taiwan
e-mail: hker@mail.chihlee.edu.tw

Y.-H. Lee
Department of Civil Engineering, Tamkang University, Taipei, 251, Taiwan
e-mail: yinghaur@mail.tku.edu.tw

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_29,
© Springer Science+Business Media B.V. 2011

into the underlying causal relationships related to developmental pattern requires studying the structure of measurements taken on different occasions [1]. Multivariate analysis of variance (MANOVA), repeated measures ANOVA, and standard multiple regression methods have been the most widely used tools for analyzing longitudinal data. Polynomial functions are usually employed to model individual growth patterns.

Classical longitudinal data analysis relies on balanced designs where each individual is measured at the same time (i.e., no missing observations). MANOVA, which imposes no constraints on residual covariance matrix, is one common approach in analyzing longitudinal data. However, an unconstrained residual covariance structure is not efficient if the residual errors indeed possess a certain structure, especially when this structure is often of interest in longitudinal studies. Repeated measures ANOVA have the assumption of sphericity. It is too restrictive for longitudinal data because such data often exhibit larger correlations between nearby measurement than between measurements that are far apart. The variance and covariance of the within-subject errors also vary over time. The sphericity assumption is inappropriate in longitudinal studies if residual errors exhibit heterogeneity and dependence.

In longitudinal studies, the focus is on determining whether subjects respond differently under different treatment conditions or at different time points. The errors in longitudinal data often exhibit heterogeneity and dependence, which call for structured covariance models. Longitudinal data typically possess a hierarchical structure that the repeated measurements are nested within an individual. While the repeated measures are the first level, the individual is the second-level unit and groups of individuals are higher level units [2]. Traditional regression analysis and repeated measures ANOVA fail to deal with these two major characteristics of longitudinal data.

Linear Mixed-Effects (LME) models are an alternative for analyzing longitudinal data. These models can be applied to data where the number and the spacing of occasions vary across individuals and the number of occasions is large. LME models can also be used for continuous time. LME models are more flexible than MANOVA in that they do not require an equal number of occasions for all individuals or even the same occasions. Moreover, varied covariance structures can be imposed on the residuals based on the nature of the data. Thus, LME models are well suited for longitudinal data that have variable occasion time, unbalanced data structure, and constrained covariance model for residual errors.

A systematic modeling approach using visual–graphical techniques and LME models was proposed and demonstrated using the original AASHO road test flexible pavement data [3]. The proposed approach including characterizing the growth patterns at both group and individual levels, identifying the important predictors and unusual subjects, choosing suitable statistical models, selecting random-effects structures, suggesting possible residuals covariance models, and examining the model-fits will be further discussed [4–7].

29.2 Methods

Hierarchical linear models allow researchers to analyze hierarchically nested data with two or more levels. A two-level hierarchical linear model consists of two submodels: individual-level (level-1) and group-level (level-2). The parameters in a group-level model specify the unknown distribution of individual-level parameters. The intercept and slopes at individual-level can be specified as random. Substituting the level-2 equations for the slopes into the level-1 model yields a linear mixed-effects (LME) model. LME models are mixed-effects models in which both fixed and random effects occur linearly in the model function [8].

In a typical hierarchical linear model, the individual is the level-1 unit in the hierarchy. An individual has a series of measurements at different time points in longitudinal studies [9]. When modeling longitudinal data, the repeated measurements are the level-1 units (i.e., a separate level below individuals). The individual is the second-level unit, and more levels can be added for possible group structures [2]. The basic model at the lowest level, also regarded as repeated-measures level, for the application of hierarchical linear model in longitudinal data can be formulated as:

$$\text{Level - 1: } \mathrm{Y}_{tj} = \beta_{0j} + \beta_{1j}c_{tj} + \beta_{2j}x_{tj} + \mathrm{r}_{tj} \tag{29.1}$$

where Y_{tj} is the measure for an individual j at time t, c_{tj} is the time variable indicating the time of measurement for this individual, x_{tj} is the time-varying covariate, and r_{tj} is the residual error term.

$$\text{Level - 2: } \begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}W_{j1} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}W_{j1} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21}W_{j1} + u_{2j} \end{aligned} \tag{29.2}$$

In this level-2 equation, W is the time-invariant covariate for this individual. After substituting level-2 equations into level-1, the combined or the linear mixed-effects model is:

$$\begin{aligned} Y_{tj} = {} & [\gamma_{00} + \gamma_{10}c_{tj} + \gamma_{20}x_{tj} + \gamma_{01}W_{j1} + \gamma_{11}W_{j1}c_{tj} + \gamma_{21}W_{j1}x_{tj}] \\ & + [u_{0j} + u_{1j}c_{tj} + u_{2j}x_{tj} + r_{tj}] \end{aligned} \tag{29.3}$$

The level-1 model is a within-individuals model and the level-2 model is a between-individuals model [10]. Note that there is no time-invariant covariate in level-2 before introducing the variable W. The variance and covariance of the u's are the variances and covariances of the random intercept and slopes. After introducing the variable W, the variance and the covariance of u's are the variance and covariance of residual intercept and slopes after partitioning out the variable W. More time-invariant variables can be added sequentially into level-2 to get different models. The reduction in variance of u's could provide an estimate of variance in intercepts and slopes accounted for by those W's [11]. This linear mixed-effects model does not require that every individual must have the same

number of observations because of possible withdrawal from study or data transmission errors.

Let $\boldsymbol{Y}_{tj}$ denotes the tth measurement on the jth individual, in which $t = 1, 2, \ldots, n_i$ measurements for subject j, and $j = 1, 2, \ldots, N$ individuals. The vector $\boldsymbol{Y}_j$ is the collection of the observations for the jth individual. A general linear mixed-effects model for individual j in longitudinal analysis can be formulated as:

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{U}_j + \mathbf{R}_j \tag{29.4}$$

where $\mathbf{X}_j$ is an $(n_j \times p)$ design matrix for the fixed effects; and $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of fixed-effect parameters. $\mathbf{Z}_j$ is an $(n_j \times r)$ design matrix for the random effects; and $\boldsymbol{U}_j$ is an $(r \times 1)$ vector of random-effect parameters assumed to be independently distributed across individuals with a normal distribution, $\boldsymbol{U}_j \sim \text{NID}(\mathbf{0}, \mathbf{T})$. The $\boldsymbol{U}_j$ vector captures the subject-specific mean effects as well as reflects the extra variability in the data. $\boldsymbol{R}_j$ is an $(n_j \times 1)$ vector for the residuals. The within errors, $\boldsymbol{R}_j$, are assumed normally distributed with mean zero and variance $\sigma^2\mathbf{W}_j$, where $\mathbf{W}_j$ (stands for "within") is a covariance matrix with a scale factor σ^2. The matrix $\mathbf{W}_j$ can be parameterized by using a few parameters and assumed to have various forms, e.g., an identity matrix or the first-order of autoregression or moving-average process [12, 13]. They are independent from individual to individual and are independent of random effects, $\boldsymbol{U}_j$.

Other choices for variance–covariance structures that involve correlated within-subject errors have been proposed. Using appropriate covariance structures can increase efficiency and produced valid standard errors. The choice among covariance depends upon data structures, subject-related theories and available computer packages. In some cases, heterogeneous error variances can be employed in the model because the variances in this model are allowed to increase or decrease with time. The assumption of common variance shared by all individuals is removed [12, 14].

LME models generally assume that level-1 residual errors are uncorrelated over time. This assumption is questionable for longitudinal data that have observations closely spaced in time. There typically exists dependence between adjacent observations. This is called serial correlation and it tends to diminish as the time between observations increases. Serial correlation is part of the error structure and if it is present, it must be part of the model for producing proper analysis [12]. If the dependent within-subject errors are permitted, the choice of the model to represent the dependence needs careful consideration. It would be preferable to incorporate as much individual-specific structure as possible before introducing a serial correlation structure into within-subject errors [15].

29.3 Data Description

The AASHO road test was a large-scale highway research project conducted near Ottawa, Illinois from 1958 to 1960, and has had by far the largest impact on the history of pavement performance analysis. The test consisted of six loops,

numbered 1–6. Each loop was a segment of a four-lane divided highway and centerlines divided the pavements into inner and outer lanes, called lane 1 and lane 2. Pavement designs varied from section to section. All sections had been subjected to almost the same number of axle load applications on any given date. Performance data was collected based on the trend of the pavement serviceability index at 2-week interval. The last day of each 2-week period was called an "index day." Index days were numbered sequentially from 1 (November 3, 1958) to 55 (November 30, 1960) [3, 7, 16].

Empirical relationships between pavement thickness, load magnitude, axle type, accumulated axle load applications, and performance trends for both flexible and rigid pavements were developed after the completion of the road test. Several combinations of certain rules, mathematical transformations, analyses of variance, graphs, and linear regression techniques were utilized in the modeling process to develop such empirical relationships. A load equivalence factor was then established to convert different configurations of load applications to standard 18-kip equivalent single-axle loads (ESAL). This ESAL concept has been adopted internationally since then. As pavement design evolves from traditional empirically based methods toward mechanistic-empirical, the ESAL concept used for traffic loads estimation is no longer adopted in the recommended Mechanistic-Empirical Pavement Design Guide (MEPDG) [17], although many researchers have argued that it is urgently in need of reconsideration [3, 18, 19].

During the road test, it was found that the damage rate was relatively low in winter but was relatively high in spring for flexible pavements. Therefore, load applications were adjusted by "seasonal weighting function" such that a better "weighted" flexible pavement equation was developed. Lee [18] has pointed out that the error variance increases when the predicted number of weighted load repetitions (W) increases. To serve the needs of predicting pavement serviceability index (PSI) after certain load applications on a given section, it is not uncommon that engineers would rearrange the original flexible pavement equation into the following form:

$$\begin{aligned} \text{PSI} &= 4.2 - 2.7 * 10^{\left[0.4+\frac{1094}{(SN+1)^{5.19}}\right]*[\log(ESAL)-9.36*\log(SN+1)+0.2]} \\ \text{SN} &= 0.44D_1 + 0.14D_2 + 0.11D_3 \end{aligned} \tag{29.5}$$

In which the regression statistics are: $R^2 = 0.212$, SEE $= 0.622$, $N = 1083$ [18]. Note that PSI ranges from 0 to 5 (0–1 for very poor; 1–2 for poor; 2–3 for fair; 3–4 for good; and 4–5 for very good conditions). D_1 is the surface thickness (in.); D_2 is the base thickness (in.); D_3 is the subbase thickness (in.).

29.4 Exploratory Analysis

Exploratory analysis is a technique to visualize the patterns of data. It is detective work of exposing data patterns relative to research interests. Exploratory analysis of longitudinal data can serve to: (a) discover as much of the information regarding

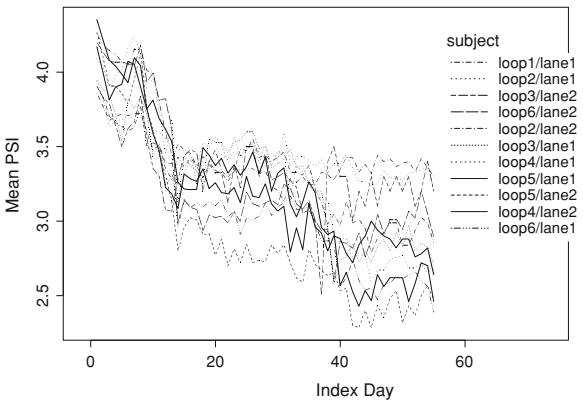


Fig. 29.1 Mean PSI for each subject (loop/lane) versus index day

raw data as possible rather than simply summarize the data; (b) highlight mean and individual growth patterns which are of potential research interest; as well as (c) identify longitudinal patterns and unusual subjects. Hence plotting individual curves to carefully examine the data should be performed first before any formal curve fitting is carried out. For the nature of this flexible pavement data, the exploratory analysis includes exploring "growth" patterns and the patterns regarding experimental conditions.

29.4.1 Exploring "Growth" Patterns

The first step, which is perhaps the best way to get a sense of a new data, is to visualize or plot the data. Most longitudinal data analyses address individual growth patterns over time. Thus, the first useful exploratory analysis is to plot the response variable against time including individual and overall mean profiles. Individual mean profiles, which summarize the aspects of response variable for each individual over time, can be used to examine the possibility of variations among individuals and to identify potential outliers. The overall mean profile summarizes some aspects of the response variable over time for all subjects and is helpful in identifying unusual time when significant differences arise.

Figure 29.1 shows the lines connecting the dependent variable (mean PSI) over time for each subject (loop/lane). Most subjects have higher mean PSIs at the beginning of the observation period, and they tend to decrease over time. The spread among the subjects is substantially smaller at the beginning than that at the end. In addition, there exist noticeable variations among subjects. The overall mean growth curve over time indicates that the overall mean PSIs are larger at the beginning and decrease over time; and the rate of deterioration is higher at the beginning than that at the end.

29.4.2 Exploring the Patterns of Experimental Conditions

In addition to time (in terms of index day), different major experimental conditions may be considered. This exploratory analysis is intended to discover the overall and individual patterns of each experimental condition and their interactions on mean PSIs. Subsequently, the patterns of mean PSIs for each subject and the patterns of overall mean PSIs on each experimental condition and their interactions over time are investigated [7]. Generally speaking, the mean PSIs for pavements with higher surface thickness are higher than those with thinner surface layer.

29.5 Linear-Mixed Effects Modeling Approach

The following proposed modeling approach is generally applicable to modeling multilevel longitudinal data with a large number of time points. Model building procedures including the selection of a preliminary mean structure, the selection of a random structure, the selection of a residual covariance structure, model reduction, and the examination of the model fit are subsequently illustrated.

29.5.1 Selecting a Preliminary Mean Structure

Covariance structures are used to model variation that cannot be explained by fixed effects and depend highly on the mean structures. The first step to model building is to remove the systematic part and remove this so that the variation can be examined. The dataset includes the following explanatory variables: thick, basethk, subasthk, uwtappl, FT. In which, thick is the surface thickness (in.); basethk is the base thickness (in.); subasthk is the subbase thickness (in.); uwtappl is the unweighted applications (millions), and FT is monthly the freeze–thaw cycles.

A model containing all main effects, and all the two-way, three-way interaction terms was first investigated. This model (called model-1) has the form:

$$\begin{aligned}\overline{\mathrm{PSI}_{ij}} = {} & \beta_{0j} + \beta_{1j}(\mathrm{thick})_{ij} + \beta_{2j}(\mathrm{basethk})_{ij} + \beta_{3j}(\mathrm{subasthk})_{ij} \\ & + \beta_{4j}(\mathrm{uwtappl})_{ij} + \beta_{5j}(\mathrm{uwtappl})_{ij}^{2} + \beta_{6j}(\mathrm{FT}) + \text{two-way} \\ & \text{interaction terms of thick, basethk, subasthk, and uwtappl} \\ & + \text{three-way interaction terms of thick, basethk, subasthk,} \\ & \text{and uwtappl} + R_{ij} \end{aligned} \tag{29.6}$$

29.5.2 Selecting a Preliminary Random Structure

The second step is to select a set of random effects in the covariance structure. An appropriately specified covariance structure is helpful in interpreting the random variation in the data, achieving the efficiency of estimation, as well as obtaining valid inferences of the parameters in the mean structure of the model. In longitudinal studies, the same subject is repeatedly measured over time. The data collected from longitudinal study is a collection of correlated data. The within-subject errors are often heteroscedastic (i.e., having unequal variance), correlated, or both.

29.5.2.1 Exploring Preliminary Random-Effects Structure

A useful tool to explore the random-effects structure is to remove the mean structure from the data and use ordinary least square (OLS) residuals to check the need for a linear mixed-effects model and decide which time-varying covariate should be included in the random structure.

The boxplot of residuals by subject corresponding to the fit of a single linear regression by using the same form of the preliminary level-1 model was conducted. This is the case when grouping structure is ignored from the hierarchy of data. Since the residuals are not centered around zero, there are considerable differences in the magnitudes of residuals among subjects. This indicates the need for subject effects, which is precisely the motivation for using linear mixed-effects model. Separate linear regression models were employed to fit each subject to explore the potential linear relationship.

To assist in selecting a set of random effects to be included in the covariance model, the plots of mean raw residuals against time and the variance of residuals against time are useful. If only random-intercepts models hold, the residual has the form, $e_{ij} = U_{0j} + R_{ij}$, in which $\boldsymbol{U}_{0j}$ is the random effect for intercepts and $\boldsymbol{R}_{ij}$ is the level-1 error. If this plot shows constant variability over time or the curves are flat, then only random intercept model is needed. If random-intercepts-and-slopes models hold, the residual has the form, $e_{ij} = U_{0j} + U_{1j}x_{1ij} + \cdots + U_{qj}x_{qij} + R_{ij}$, where $\boldsymbol{U}_{qj}$ is the random effect for the qth slope. In the case of random-intercepts-and-slopes model, the plot would show the variability varies over time or there are some unexplained systematic structures in the model. One or more random effects, additional to random intercept, have to be added.

29.5.2.2 Selecting a Variance–Covariance Structure for Random Effects

Three possible variance–covariance structures including general positive definite (unstructured), diagonal, and block-diagonal based on different assumptions [8] were investigated. General positive-definite is a general covariance matrix

Table 29.1 Model comparison using three variance–covariance structures

Model	df	AIC	BIC	logLik	Test	L. ratio	*p*-Value
(1) Unstr	29	12910.29	13117.74	−6426.14			
(2) Diag	22	13056.52	13213.90	−6506.26	1 vs 2	160.234	<0.0001
(3) Bk-diag	21	13060.14	13210.37	−6509.07	2 vs 3	5.621	0.0177

parameterized directly in terms of variances and covariances. Diagonal covariance structure is used when random-effects are assumed independent. Block-diagonal matrix is employed when it is assumed that different sets of random effects have different variances. Table 29.1 displays the model comparison of these three models. The unstructured model has the smallest absolute value of log-likelihood among them. The likelihood ratio test for unstructured model versus diagonal model is 160.23 with *p*-value less than 0.0001. Thus, unstructured variance–covariance model will be used hereafter.

The random effects of the preliminary level-2 model include intercept, uwtappl, quadratic term of uwtappl, and FT. The variance–covariance structure is a general positive-definite matrix. Putting the preliminary level-1 and level-2 models together, the preliminary linear-mixed-effects model is then:

$$\begin{aligned}\overline{\mathrm{PSI}_{ij}} = {} & \gamma_{00} + \gamma_{10}(\mathrm{thick})_{ij} + \gamma_{20}(\mathrm{basethk})_{ij} + \gamma_{30}(\mathrm{subasthk})_{ij} \\ & + \gamma_{40}(\mathrm{uwtappl})_{ij} + \gamma_{50}(\mathrm{uwtappl})^2_{ij} + \gamma_{60}(\mathrm{FT})_{ij} + \gamma_{70}(\mathrm{thick} \\ & * \mathrm{basethk})_{ij} + \gamma_{80}(\mathrm{thick} * \mathrm{subasthk})_{ij} + \gamma_{90}(\mathrm{basethk} \\ & * \mathrm{uwtappl})_{ij} + \gamma_{100}(\mathrm{subasthk} * \mathrm{uwtappl})_{ij} + \gamma_{110}(\mathrm{basethk}* \\ & \mathrm{subasthk} * \mathrm{uwtappl})_{ij} + \gamma_{120}(\mathrm{thick} * \mathrm{basethk} * \mathrm{subasthk} * \mathrm{uwtappl})_{ij} \\ & + U_{0j} + U_{4j}(\mathrm{uwtappl})_{ij} + U_{5j}(\mathrm{uwtappl})^2_{ij} + U_{6j}(FT)_{ij} + R_{ij}\end{aligned} \tag{29.7}$$

29.5.3 Selecting a Residual Covariance Structure

The absolute value of log-likelihood for this heteroscedastic model is 6273.29. The need of heteroscedastic model can be formally checked by using the likelihood ratio test [7]. The small *p*-value indicates that the heteroscedastic model explains the data significantly better than homoscedastic model.

Correlation structures are used to model dependence among the within-subject errors. Autoregressive model with order of 1, called AR(1), is the simplest and one of the most useful models [8]. The autocorrelation function (ACF), which begins autocorrelation at lag 1 and then declines geometrically, for AR(1) is particularly simple. Autocorrelation functions for autoregressive model of order greater than one are typically oscillating or sinusoidal functions and tend to damp out with increasing lag [20].

Thus, AR(1) may be adequate to model the dependency of the within-subject errors. The absolute value of log-likelihood for this heteroscedastic AR(1) model is 6207.24. The estimated single correlation parameter ϕ is 0.125. The heteroscedastic model (corresponding to $\phi = 0$) is nested within the heteroscedastic AR(1) model.

Likewise, the need of heteroscedastic AR(1) model can be checked using likelihood ratio test [7]. The small p-value indicates that the heteroscedastic AR(1) model explains the data significantly better than heteroscedastic model, suggesting that within-group serial correlation is present in the data.

29.5.4 Model Reduction

After specifying the within-subject error structure, the next step is to check whether the random-effects can be simplified. It is also desirable to reduce the number of parameters in fixed effects in order to achieve a parsimonious model that can well represent the data. A likelihood ratio test statistic, whose sampling distribution is a mixture of two chi-squared distributions, is used to test the need for random-effects. The p-value is determined by equal weight of the p-values of a mixture of two chi-squared distributions. To assess the significance of the terms in the fixed effects, conditional t-tests are used.

29.5.4.1 Reduction of Random Effects

The matrix of known covariates should not have polynomial effect if not all hierarchically inferior terms are included [21]. The same rule applies to interaction terms. Hence, significance tests for higher-order random effects should be performed first. The random effects included in the preliminary random-effects structure are: intercept, uwtappl, uwtappl^2, and FT. The models and the associated maximum log-likelihood values are compared [7]. The small p-value indicates that the preliminary random-effects structure explains the data significantly better than the others. Thus, no reduction of random effects is needed.

29.5.4.2 Reduction of Fixed Effects

An adequate and appropriately specified random-effects structure implies efficient model-based inferences for the fixed effects. When considering the reduction of fixed effects, one model is nested within the other model and the random-effects structures are the same for the full and the reduced models. Likelihood ratio tests are appropriate for the model comparison. The parameter estimates, estimated standard errors, t-statistics and p-values for the fixed effects of the heteroscedastic AR(1) model are revisited. The heteroscedastic AR(1) model can be reduced to a

Table 29.2 Proposed preliminary LME model

	Intercept	uwtappl	uwtappl^2	FT	Residual
Random effects					
Standard deviation	0.170	1.679	0.765	0.00722	0.448

Parameter	Value	Std. error	DF	*t*-Value	*p*-Value
Fixed effects					
(Intercept)	2.4969	0.0703	9423	35.51	<0.0001
thick	0.2629	0.0122	9423	21.48	<0.0001
basethk	0.0590	0.0066	9423	8.91	<0.0001
subasthk	0.0386	0.0041	9423	9.37	<0.0001
uwtappl	−3.6191	0.5254	9423	−6.89	<0.0001
uwtappl^2	1.1524	0.2481	9423	4.65	<0.0001
FT	0.0148	0.0023	9423	6.39	<0.0001
thick*basethk	−0.0062	0.0016	9423	−3.81	<0.0001
thick*subasthk	−0.0082	0.0010	9423	−8.07	<0.0001
basethk*uwtappl	0.1275	0.0172	9423	7.40	<0.0001
subasthk*uwtappl	0.1355	0.0181	9423	7.50	<0.0001
thick*basethk*uwtappl	−0.0155	0.0045	9423	−3.43	0.0006
thick*subasthk*uwtappl	−0.0077	0.0036	9423	−2.16	0.0307
basethk*subasthk*uwtappl	−0.0291	0.0029	9423	−9.87	<0.0001
thick*basethk*subasthk*uwtappl	0.0073	0.0006	9423	11.53	<0.0001

Note Model fit: AIC = 12481.77, BIC = 12710.69, logLik = −6208.89. Correlation structure: AR(1); parameter estimate(s): Phi = 0.126. Variance function structure: for different standard deviations per stratum (thick = 2, 1, 3, 4, 5, 6 in.), the parameter estimates are: 1, 1.479, 0.935, 1.199, 0.982, 0.959

more parsimonious model due to the existence of some insignificant parameter estimates. The reduction of fixed effects starts with removing the parameters with largest *p*-values, insignificant terms, and combining the parameters not changing significantly. These processes are repeated until no important terms have been left out of the model.

29.5.5 Proposed Preliminary LME Model

The final proposed preliminary linear mixed-effects model is listed in Table 29.2. The fixed-effects structures of the proposed model contain significant treatment effects for thick, basethk, subasthk, uwtappl, uwtappl^2, FT, and several other two-, three-, and four-way interaction terms. The positive parameter estimates for thick, basethk, and subasthk indicates that higher mean PSI values tend to occur on thicker pavements. The parameter estimate of uwtappl is negative indicating that lower PSI values for higher load applications.

Furthermore, the preliminary LME model also indicates that: The standard error for the pavements with surface thickness of 1 in. or 4 in. is about 48 or 20% higher

than those with surface thickness of 2 in., respectively. There exists dependency in within-subject errors. The estimated single correlation parameter for the AR(1) model is 0.126.

29.5.6 Examination of the Model Fit

A plot of the population predictions (fixed), within-group predictions (Subject), and observed values versus time for the proposed preliminary LME model by subjects. Population predictions are obtained by setting random-effects to zero whereas within-group predictions use estimated random effects [7]. The prediction line of the within-group predictions follows the observed values more closely indicating the proposed LME model provides better explanation to the data.

29.6 Conclusions

A systematic modeling approach using visual-graphical techniques and LME models which is generally applicable to modeling multilevel longitudinal data with a large number of time points was proposed in this paper. The original AASHO road test flexible pavement data was used to illustrate the proposed modeling approach.

Exploratory analysis of the data indicated that most subjects (loop/lane) have higher mean PSIs at the beginning of the observation period, and they tend to decrease over time. The spread among the subjects is substantially smaller at the beginning than that at the end. In addition, there exist noticeable variations among subjects.

A preliminary LME model for PSI prediction was developed. The positive parameter estimates for thick, basethk, and subasthk indicates that higher mean PSI values tend to occur on thicker pavements. The parameter estimate of uwtappl is negative indicating that lower PSI values for higher load applications. The prediction line of the within-group predictions (Subject) follows the observed values more closely than that of the population predictions (fixed) indicating the proposed LME model provides better explanation to the data.

References

1. Goldstein H (1979) The design and analysis of longitudinal studies. Academic Press, Inc, New York
2. Hox JJ (2000) Multilevel analysis of grouped and longitudinal data. In: Little TD, Schnabel KU, Baumert J (eds) Modeling longitudinal and multilevel data: practical issues, applied approaches and specific examples. Lawrence Erlbaum Associates, Mahwah, pp 15–32
3. Highway Research Board (1962) The AASHO road test, report 5, pavement research, special report 61E. National Research Council, Washington

4. Ker HW (2002) Application of regression spline in multilevel longitudinal modeling. Doctoral Dissertation, University of Illinois, Urbana
5. Lee YH, Ker HW (2008) Reevaluation and application of the AASHTO mechanistic-empirical pavement design guide, phase I, summary report, NSC96-2211-E-032-036. National Science Council, Taipei City (In Chinese)
6. Lee YH, Ker HW (2009) Reevaluation and application of the AASHTO mechanistic-empirical pavement design guide, phase II, NSC97-2221-E-032-034, summary report. National Science Council, Taipei City (In Chinese)
7. Ker HW, Lee YH (2010) Preliminary analysis of AASHO road test flexible pavement data using linear mixed effects models. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July, London, UK, pp 260–266
8. Pinherio JC, Bates DM (2000) Mixed-effects models in S and S-plus. Springer, New York
9. Laird NM, Ware JH (1982) Random effects models for longitudinal data. Biometrics 38: 963–974
10. Anderson CJ (2001) Model building. http://www.ed.uiuc.edu/courses/edpsy490ck
11. MacCallum RC, Kim C (2000) Modeling multivariate change. In: Little TD, Schnabel KU, Baumert J (eds) Modeling longitudinal and multilevel data: practical issues, applied approaches and specific examples. Lawrence Erlbaum Associates, NJ, pp 51–68
12. Jones RH (1993) Longitudinal data with serial correlation: a state-space approach. Chapman & Hall, London
13. Vonesh EF, Chinchilli VM (1997) Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker, Inc, New York
14. Carlin BP, Louis TA (1996) Bayes and empirical Bayes methods for data analysis. Chapman & Hall, London
15. Goldstein H, Healy MJR, Rasbash J (1994) Multilevel time series models with application to repeated measures data. Stat Med 13:1643–1655
16. Huang YH (2004) Pavement analysis and design, 2nd edn. Pearson Education, Inc., Upper Saddle River
17. ARA, Inc (2004) ERES consultants division, guide for mechanistic-empirical design of new and rehabilitated pavement structure. NCHRP 1–37A report. Transportation Research Board, National Research Council, Washington
18. Lee YH (1993) Development of pavement prediction models. Doctoral dissertation, University of Illinois, Urbana
19. Ker HW, Lee YH, Wu PH (2008) Development of fatigue cracking performance prediction models for flexible pavements using LTPP database. J Transp Eng ASCE 134(11):477–482
20. Pindyck RS, Rubinfeld DL (1998) Econometric models and economic forecasts, 4th edn. McGraw-Hill, Inc, New York
21. Morrell CH, Pearson JD, Brant LJ (1997) Linear transformations of linear mixed-effects models. Am Stat 51:338–343

Chapter 30
Chi-Squared, Yule's Q and Likelihood Ratios in Tabular Audiology Data

Muhammad Naveed Anwar, Michael P. Oakes and Ken McGarry

Abstract In this chapter, we have used the chi-squared test and Yule's Q measure to discover associations in tables of patient audiology data. These records are examples of heterogeneous medical records, since they contain audiograms, textual notes and typical relational fields. In our first experiment we used the chi-squared measure to discover associations between the different fields of audiology data such as patient gender and patient age with diagnosis and the type of hearing aid worn. Then, in our second experiment we used Yule's Q to discover the strength and direction of the significant associations found by the chi-squared measure. Finally, we examined the likelihood ratio used in Bayesian evidence evaluation. We discuss our findings in the context of producing an audiology decision support system.

M. N. Anwar (✉) · M. P. Oakes
Department of Computing, Engineering & Technology, University of Sunderland, Sunderland, UK
e-mail: Naveed.Anwar@sunderland.ac.uk

M. P. Oakes
e-mail: Michael.Oakes@sunderland.ac.uk

K. McGarry
Department of Pharmacy, Health and Well-Being, University of Sunderland, Sunderland, UK
e-mail: Ken.Mcgarry@sunderland.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_30,
© Springer Science+Business Media B.V. 2011

30.1 Introduction

Association measures can be used to measure the strength of relationship between the variables in medical data. Discovering associations in medical data has an important role in predicting the patient's risk of certain diseases. Early detection of any disease can save time, money and painful procedures [1]. In our work we are looking for significant associations in heterogeneous audiology data with the ultimate aim of looking for factors influencing which patients would most benefit from being fitted with a hearing aid.

Support and confidence are measures of the interestingness of associations between variables [2, 3]. They show the usefulness and certainty of discovered associations. Strong associations are not always interesting, because support and confidence do not filter out uninteresting associations [4]. Thus, to overcome this problem a correlation measure is augmented to support and confidence. One of the correlation measures popularly used in the medical domain is chi-squared (χ^2).

In Sect. 30.2 we describe our database of audiology data. We first use the chi-squared measure to discover significant associations in our data, as described in Sect. 30.3. We then use Yule's Q measure to discover the strength of each of our significant associations, as described in Sect. 30.4. In Sect. 30.5, we briefly describe our findings for the support and confidence for each of the significant associations. In Sect. 30.6, we use Bayesian likelihood ratios to find associations between words in the comments fields and the type of hearing aid fitted. We draw our conclusions in Sect. 30.7.

30.2 Audiology Data

In this study, we have made use of audiology data collected at the hearing aid outpatient clinic at James Cook University Hospital in Middlesbrough, England, UK. The data consists of about 180,000 individual records covering about 23,000 audiology patients. The data in the records is heterogeneous, consisting of the following fields:

1 Audiograms, which are the graphs of hearing ability at different frequencies (pitches).
2 Structured data: gender, date of birth, diagnosis and hearing aid type, as stored in a typical database, e.g. |M|, |09-05-1958|, |TINNITUS|, |BE18|.
3 Textual notes: specific observations made about each patient, such as |HEARING TODAY NEAR NORMAL—USE AID ONLY IF NECESSARY|.

In general, these audiology records represent all types of medical records because they involve both structured and unstructured data.

30.3 Discovery of Associations with the Chi-Squared Test Tables

The chi-squared test is a simple way to provide estimates of quantities of interest and related confidence intervals [5]. It is a measure of associations between variables (such as the fields of the tables in a relational database) where the variables are nominal and related to each other [6]. The Chi-squared test is popular in the medical domain because of its simplicity. It has been used in pharmacology to classify text according to subtopics [7]. The resulting chi-squared value is a measure of the differences between a set of observed and expected frequencies within a population, and is given by the formula [5]:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where r is the number of unique terms in a particular field of the patient records such as diagnosis or hearing aid type, corresponding to rows in Table 30.1. c is the number of categories in the data (such as age or gender) corresponding to columns in Table 30.1.

Table 30.1 is produced for two diagnoses occurring in the hearing diagnosis field. For example, if 535 of the hearing diagnosis fields of the records of patients 'Aged $\leq$ 54' years contained the diagnosis 'tinnitus', we would record a value of 535 for that term being associated with that category. These values were the 'observed' values, denoted O_{ij} in the formula above. The corresponding 'expected' values E_{ij} were found by the formula:

$$\text{Row total} \times \text{Column total}/\text{Grand Total}$$

The row total for 'tinnitus' diagnosis is the total number of times the 'tinnitus' diagnosis was assigned to patients in both age categories = 535 + 592 = 1127. The column total for 'Age $\leq$ 54' is the total number of patients in that age group over all two diagnoses = 702. The grand total is the total number of patient records in the study = 1364. Thus the 'expected' number of patients diagnosed with 'tinnitus' in the 'Age $\leq$ 54' group was 1127 * 702/1364 = 580.02. The significance of this is that the expected value is greater than the observed value, suggesting that there is a negative degree of association between the 'tinnitus'

Table 30.1 Observed and expected frequencies for diagnosis

Diagnosis	Age $\leq$ 54	Age > 54	Row total
Not-tinnitus	167 (121.98) [2027.24]	70 (115.02) [2027.24]	237
Tinnitus	535 (580.02) [2027.24]	592 (546.98) [2027.24]	1127
Column total	702	662	1364

Expected frequencies are in ( ); (observed frequency − expected frequency)^2 are in [ ]

diagnosis and the category 'Age $\leq$ 54'. The remainder of the test is then performed to discover if this association is statistically significant. Since we were in effect performing many individual statistical tests, it was necessary to use the Bonferroni correction [5] to control the rate of Type I errors where a pair of variables spuriously appear to be associated. For example, for us to be 99.9% confident that a particular keyword was typical of a particular category, the corresponding significance level of 0.001 had to be divided by the number of simultaneous tests, i.e. the number of unique words times the number of categories. In the case of words in the text fields, this gave a corrected significance level of $0.001/(2 * 2) = 0.00025$. Using West's chi-squared calculator [8], for significance at the 0.001 level with one degree of freedom, we obtained a chi-squared threshold of 13.41. Thus each word associated with a category with a chi-squared value of more than 13.41 was taken to be significantly associated with that category at the 0.001 level.

The overall chi-squared values for the relationship between the test variables age and gender with hearing aid type (behind the ear—BTE/in the ear—ITE) are shown in Table 30.2. The overall chi-squared value for the relationship between the words in the comments text and hearing aid type was calculated by summing the chi-squared values of all possible text word—BTE/ITE right aid pairs, and is also shown in Table 30.2. This data shows, with 99.9% confidence, that these text words were not randomly distributed, but some text words are probably associated with hearing aid type. Similarly the associations of each of the variables (age, comments text, gender and hearing aid type) with tinnitus diagnosis are shown in Table 30.3. Here we see that there are significant associations between age, comments text, and BTE/ITE right aid with a diagnosis of tinnitus, but there are no significant associations between gender and tinnitus diagnosis.

Table 30.2 Overall χ^2 with BTE/ITE right aid

Fields	Overall χ^2	Degrees of freedom (df)	*P*
Age	10.53	1	<0.001
Comments text	5421.84	663	<0.001
Gender	33.68	1	<0.001

Table 30.3 Overall χ^2 with tinnitus diagnosis

Fields	Overall χ^2	Degrees of freedom (df)	*P*
Age	41.45	1	<0.001
Comments text	492.26	60	<0.001
Gender	0.18	1	=0.6714
BTE/ITE raid	31.75	1	<0.001

To use the chi-squared test the expected frequency values must be all at least 1, and most should exceed 5 [9]. To be on the safe side, we insisted that for each word, all the expected values should be at least 5, so all words failing this test were grouped into a single class called 'OTHERS'.

Keywords associated with categories with 95% confidence were deemed typical of those categories if $O > E$, otherwise they were deemed atypical. The keywords most typical and atypical of the four categories (hearing aid type, age, tinnitus and gender) are shown in Tables 30.4 and 30.5. A 'keyword' could either be a category type (where * denotes a diagnosis category, and *** denotes a hearing aid category), or a word from the free-text comments field (denoted **). The discovered associations seem intuitively reasonable. For example, it appears that the patients with 'Age $\leq$ 54' tend not to have tinnitus, and patients not having tinnitus had a problem of wax and were using BTE hearing aids. The words tinnitus (ringing in the ears) and masker (a machine for producing white noise to drown out tinnitus) were atypical of this category. It was found that males tended more to use ITE hearing aids and females tended more to use BTE hearing aids. The hearing aid types associated with BTE were those with high gain and had changes made to the ear mould. Similarly, ITE hearing aid types used lacquer, vents, required reshelling of ear impressions, had changes made to the hearing aid, were reviewed and the wearers were making progress.

For these experiments, we used all the records available in the database for each field under study, keeping the criterion that none of the field values should be empty. In Table 30.4, 70 was calculated as the median age of the BTE/ITE right aid group and in Table 30.5, 54 was the median age of the records with a not-tinnitus or tinnitus diagnosis. In Tables 30.4 and 30.5 some keywords in the comments text were abbreviations such as 'reshel' for 'reshell' and 'fta' for 'failed to attend appointment'. 'Tinnitus' appears as 'tinnitu' in the tables, since all the text was passed through Porter's stemmer [10] for the removal of grammatical endings.

Table 30.4 Categories with positive and negative keywords in records with BTE/ITE right aid

	Positive keywords	Negative keywords
Age $\leq$ 70	**Not found*	**Not found*
Age $>$ 70	**Not found*	**Not found*
BTE	**mould, be34, map, gp, 92, audio, inf, be52, ref, staff, reqd, be36, contact	**fta, reshel, appt, it, nn, nfa, 2001, rev, lacquer, hn, km, imp, review, 2000
ITE	**fta, reshel, appt, it, nn, nfa, 2001, rev, lacquer, hn, km, imp, review, 2000, nh, vent, progress, aid, dt, taken	**mould, be34, map, gp, 92, audio, inf, be52, ref, staff, reqd, be36, contact, tri, n, order
Male	***ITE	***BTE
Female	***BTE	***ITE

Table 30.5 Categories with positive and negative keywords in records with a tinnitus/not-tinnitus diagnosis

	Positive keywords	Negative keywords
Age $\leq$ 54	*Not-tinnitus	*_Not found_
Age $>$ 54	*_Not found_	*Not-tinnitus
Not-tinnitus	**OTHERS, lost, ear, wax, L, aid ***BTE	**masker, tinnitu, rev, help, appt, 2001, 2000, counsel, ok, further, progress, fta ***ITE
Tinnitus	**masker, tinnitu ***_Not found_	**OTHERS ***_Not found_
Male	***_Not found_	***_Not found_
Female	***_Not found_	***_Not found_

30.4 Measures of Association in Categorical Data

Yule's Q is a measure to find the strength of association between categorical variables. Unlike the chi-squared test, which tells us how certain we can be that a relationship between two variables exists, Yule's Q gives both the strength and direction of that relationship [6]. In the following 2 $\times$ 2 table,

	Present	Absent
Present	A	B
Absent	C	D

Yule's Q is given by

$$Q = \frac{AD - BC}{AD + BC} \tag{2}$$

where A, B, C and D are the observed quantities in each cell. Yule's Q is in the range -1 to $+1$, where the sign indicates the direction of the relationship and the absolute value indicates the strength of the relationship. Yule's Q does not distinguish complete associations (where one of the cell values $= 0$) and absolute relationships (where two diagonally opposite cell values are both zero), and is only suitable for 2 $\times$ 2 tables.

In Tables 30.6, 30.7, 30.8, and 30.9 Yule's Q values for age with comment text, diagnosis, hearing aid type, and mould are given. Similarly, in the Table 30.10, 30.11, and 30.12 Yule's Q values for gender with comment text, hearing aid type and mould are given. '(P)' and '(A)', stand for present and absent.

In Table 30.6, a Yule's Q value of 0.75 shows that there is a positive association between the keyword 'progress' and the category 'Age $\leq$ 70', which can be restated as a negative association between the keyword 'progress' and the category 'Age $>$ 70'.

In Table 30.7, for 'diagnosis' there is an absolute association between 'familial' and 'Age $\leq$ 54', resulting in a Yule's Q value of 1. This should be viewed in comparison to the chi-squared value for the same association, 17.20 ($P < 0.001$), showing both that the association is very strong and that we can be highly confident that it exists. The presence of this association shows that a higher proportion of younger people report to the hearing aid clinic with familial (inherited) deafness than older people.

Familial deafness is relatively rare but can affect any age group, while 'OTHERS' would include 'old-age deafness' (presbycusis) which is relatively common, but obviously restricted to older patients. However, in Table 30.9, Yule's Q for 'V2' is 0.18, which shows only a weak association between mould and 'Age $\leq$ 70', while the chi-squared value for the same association of 30.25 ($P < 0.001$), showed that it is highly likely that the association exists. In

Table 30.6 Yule's Q for comment text and age

Comment text	Age $\leq$ 70 (P)	Age $>$ 70 (P)	Age $\leq$ 70 (A)	Age $>$ 70 (A)	Yule's Q
Progress	93	13	46833	45555	0.75
Dna	105	20	46821	45548	0.67
Masker	565	126	46361	45442	0.63
Tinnitus	385	123	46541	45445	0.51
Help	222	84	46704	45484	0.44
Counsel	191	80	46735	45488	0.40
2000	288	125	46638	45443	0.38
Fta	542	332	46384	45236	0.23
Gp	370	615	46162	55060	−0.16
Wax	341	601	46191	55074	−0.19
Ref	248	487	46284	55188	−0.24
Contact	37	129	46495	55546	−0.49
Insert	23	102	46509	55573	−0.58
Reqd	15	111	46517	55564	−0.72
Cic	10	76	46522	55599	−0.73
Staff	17	132	46515	55543	−0.73
Map	15	125	46517	55550	−0.75
Dv	29	245	46503	55430	−0.75
Reinstruct	8	68	46524	55607	−0.75

Table 30.7 Yule's Q for diagnosis and age

Diagnosis	Age $\leq$ 54 (P)	Age $>$ 54 (P)	Age $\leq$ 54 (A)	Age $>$ 54 (A)	Yule's Q
Familial	18	0	684	662	1.00
OTHERS	113	44	589	618	0.46

Table 30.8 Yule's Q for hearing aid type and age

Hearing aid type	Age ≤ 70 (P)	Age > 70 (P)	Age ≤ 70 (A)	Age > 70 (A)	Yule's Q
PFPPCL	42	1	11105	10899	0.95
PPCL	78	5	11069	10895	0.88
BE101	44	4	11103	10896	0.83
PPC2	53	6	11094	10894	0.79
ITENL	123	35	11024	10865	0.55
OTHERS	103	37	11044	10863	0.46
ITEHH	536	317	10611	10583	0.26
–	4668	3947	6479	6953	0.12
BE34	640	882	10507	10018	−0.18
ITENH	403	592	10744	10308	−0.21
ITENN	683	1063	10464	9837	−0.25
BE36	97	203	11050	10697	−0.37

Table 30.9 Yule's Q for mould and age

Mould	Age ≤ 70 (P)	Age > 70 (P)	Age ≤ 70 (A)	Age > 70 (A)	Yule's Q
N8	261	94	10873	10805	0.47
SIL	255	101	10879	10798	0.43
V2	575	397	10559	10502	0.18
2107V1	601	913	10533	9986	−0.23

Table 30.10 Yule's Q for comment text and gender

Comment text	M (P)	F (P)	M (A)	F (A)	Yule's Q
He	67	2	46465	55673	0.95
Wife	44	2	46488	55673	0.93
Dv	80	254	46452	55421	−0.45

Table 30.11 Yule's Q for hearing aid type and gender

Hearing aid type	M (P)	F (P)	M (A)	F (A)	Yule's Q
ITEHH	665	201	11080	12467	0.58
ITENH	725	295	11020	12373	0.47
ITEHN	1280	1732	10465	10936	−0.13
ITENN	734	1038	11011	11630	−0.14

Table 30.11, Yule's Q for 'ITEHN' (a type of hearing aid worn inside the ear) is −0.13, which shows a weak negative association between 'ITEHN' and 'male', or in other words, a weak positive association between 'ITEHN' and 'female'. In comparison, the chi-squared value for the same association of 43.36 ($P < 0.001$), showed that we can be highly confident that the relationship exists. These results show the complementary nature of the chi-squared and Yule's Q results: in all three cases the chi-squared value was highly significant, suggesting

Table 30.12 Yule's Q for mould and gender

Mould	M (P)	F (P)	M (A)	F (A)	Yule's Q
IROS	80	24	11671	12644	0.57
V2	640	342	11111	12326	0.35
N8	253	141	11498	12527	0.32

that the relationship was highly likely to exist, while Yule's Q showed the strength (strong in the first case, weak in the others) and the direction (positive in the first two cases, negative in the third) of the relationship differed among the three cases.

30.5 Support and Confidence for Associations

We examined two measures of association commonly used in market basket analysis, support and confidence [4], for all relations between age and diagnosis, and gender and diagnosis. We were unable to find many rules with high support and confidence due to the very high proportion of one type of diagnosis ('tinnitus') in the records. However, we feel that given an audiology database where a diagnosis was routinely recorded for every patient, more rules in the form $A \Rightarrow B$ (A implies B) would be found. Our results are given in [11].

30.6 Likelihood Ratios for Associated Keywords

In Bayesian Evidence Evaluation [6], the value of a piece of evidence may be expressed as a likelihood ratio (LR), as follows:

$$\mathrm{LR} = \Pr(E/H) / \Pr(E/\bar{H})$$

For example, our hypothesis (H) might be that a patient should be fitted with a BTE hearing aid as opposed to an ITE hearing aid. E is a piece of evidence such as the word 'tube' appearing in the patient's comments field of the database. $\Pr(E/H)$ is then the probability of seeing this evidence given that the hypothesis is true. Of all the 34394 records where a patient was given a BTE aid, 29 of them contained the word 'tube', so in this case $\Pr(E/H) = 29/34394 = 0.000843$. $\Pr(E/\bar{H})$ is the probability of seeing the word 'tube' when the hypothesis is not true. Of all the 29455 records where a patient was given an ITE aid, only 2 of them contained the word 'tube', so here $\Pr(E/\bar{H})$ was $2/29455 = 0.0000679$. This gives an LR of $0.000843/0.0000675 = 12.41$. Using Evett et al.'s [12] scale of verbal equivalences of the LR, an LR in the range 10–100 indicates moderate support for the hypothesis. LRs in the range 0.1–10 indicate only limited support either way, while an LR in the range 0.01 to 0.1 would indicate moderate support for the complementary hypothesis. The words giving the highest and lowest LR values

Table 30.13 Likelihood ratios for comments text and BTE/ITE right aids

Word	BTE	ITE	LR
Adequ	14	0	NA
Audiometer	10	0	NA
Be10	18	0	NA
Be201	18	0	NA
Be301	13	0	NA
Be37	12	0	NA
Be51	13	0	NA
Hac	11	0	NA
Temporari	11	0	NA
Therapy	13	0	NA
Be52	68	2	29.11
Be53	26	1	22.26
Be36	57	3	16.27
Be54	35	2	14.98
Retub	34	2	14.55
Seri	16	1	13.70
Cwc	15	1	12.84
Tube	29	2	12.41
Couldn't	14	1	11.99
Orig	14	1	11.99
"map	13	1	11.13
Map	116	9	11.03
E	12	1	10.27
Hn	8	77	0.09
Progress	4	39	0.09
Readi	1	10	0.09
Concertina	1	11	0.08
Unless	1	11	0.08
Coat	1	13	0.07
Cap	1	15	0.06
Vc	1	15	0.06
Hnv1	1	17	0.05
Hh	1	20	0.04
Reshel	6	136	0.04
Lacquer	2	65	0.03
Facepl	0	15	0
Window	0	16	0
Total	34394	29445	

with respect to a BTE fitting as opposed to an ITE fitting are shown in Table 30.13, where NA indicates division by zero as the word never appeared in records for patients fitted with an ITE hearing aid. All words which were used in the chi-squared analysis (since their expected values were all 5 or more) were also considered for this analysis.

LR values are useful for the combination of evidence. Using the evidence that the text comments field contains 'lacquer', 'reshell' and 'progress', we can

estimate the likelihood of the patient requiring a BTE hearing aid by iteratively using the relationship 'posterior odds = LR × prior odds'. Initially we obtain a prior odds (Pr(BTE)/Pr(ITE)) from a large sample or manufacturer's data. Using the column totals in Table 30.13, the prior odds in favour of a BTE aid before any other evidence has been taken into account would be 34394/29445 = 1.168 to 1. Taking the first piece of evidence (the presence of the word 'lacquer' into account), the posterior odds are 0.03 × 1.168 = 0.035. This posterior odds value now becomes the prior odds for the second iteration. The LR for 'reshell' is 0.04, so the posterior odds become 0.04 × 0.035 = 0.0014.

This posterior odds value now becomes the prior odds for the third iteration. The LR for 'progress' is 0.09, so the final posterior odds become 0.09 × 0.0014 = 0.000126. Since these posterior odds are much less than 1, it is much more likely that the patient should be fitted with an ITE hearing aid. This simple example shows the basis by which a Bayesian decision support system which returns the more suitable type of hearing aid could be constructed.

30.7 Conclusion

In this work we have discovered typical and atypical words related to different fields of audiology data, by first using the chi-squared measure to show which relations most probably exist, then using Yule's Q measure of association to find the strength and direction of those relations. The Likelihood Ratio, also based on the contingency table, provides a means whereby all the words in the comments field can be taken into account in a Bayesian decision support system for audiologists. We are currently working on the development of a Logistic Regression model, where the overall value log(Pr(BTE)/Pr(ITE)) will be a linear combination of the presence or absence of each of the discovered associated variables described in this chapter. Analogous reasoning will be used for models to predict whether or not a patient should be given a tinnitus masker, and whether or not he or she would benefit from a hearing aid fitting.

Rules found by data mining should not only be accurate and comprehensible, but also 'surprising'. McGarry presents a taxonomy of 'interestingness' measures whereby the value of discovered rules may be evaluated [13]. In this chapter, we have looked at objective interestingness criteria, such as the statistical significance of the discovered rules, but we have not yet considered subjective criteria such as unexpectedness and novelty. These require comparing machine-derived rules with the prior expectations of domain experts. A very important subjective criterion is 'actionability', which includes such considerations as impact: will the discovered rules lead to any changes in current audiological practice?

Acknowledgments We wish to thank Maurice Hawthorne, Graham Clarke and Martin Sandford at the Ear, Nose and Throat Clinic at James Cook University Hospital in Middlesbrough, England, UK, for making the large set of audiology records available to us.

References

1. Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Syst Appl, Elsevier Science Ltd 17:223–232
2. Bramer M (2007) Principles of data mining. Springer, London, pp 187–218
3. Ordonez C, Ezquerra N, Santana CA (2006) Constraining and summarizing association rules in medical data. In: Cercone N et al (eds) Knowledge and information systems. Springer, New York, pp 259–283
4. Han J, Kamber M (2006) Data mining concepts and techniques, 2nd edn. Morgan Kaufmann Publishers, San Diego, pp 227–272
5. Altman DG (1991) Practical statistics for medical research. Chapman & Hall, London, pp 241–248, 211, 271
6. Lucy D (2005) Introduction to statistics for forensic scientists. Wiley, Chichester, pp 45–52,112–114,133–136
7. Oakes M, Gaizauskas R, Fowkes H et al (2001) Comparison between a method based on the chi-square test and a support vector machine for document classification. In: Proceedings of ACM SIGIR, New Orleans, pp 440–441
8. Chi-square calculator (2010). http://www.stat.tamu.edu/~west/applets/chisqdemo.html
9. Agresti A (2002) Categorical data analysis, 2nd ed. Wiley series in probability and statistics. Wiley, New York, p 80
10. Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137
11. Anwar MN, Oakes MP, McGarry K (2010) Chi-squared and associations in tabular audiology data. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, London, UK, vol 1, pp 346–351
12. Evett IW, Jackson G, Lambert JA, McCrossan S (2000) The impact of the principles of evidence interpretation and the structure and content of statements. Sci Justice 40:233–239
13. McGarry K (2005) A survey of interestingness measures for knowledge discovery. Knowl Eng Rev J 20(1):39–61

Chapter 31
Optimising Order Splitting and Execution with Fuzzy Logic Momentum Analysis

Abdalla Kablan and Wing Lon Ng

Abstract This study proposes a new framework for high frequency trading using a fuzzy logic based momentum analysis system. An order placement strategy will be developed and optimised with adaptive neuro fuzzy inference in order to analyse the current "momentum" in the time series and to identify the current market condition which will then be used to decide the dynamic participation rate given the current traded volume. The system was applied to trading of financial stocks, and tested against the standard volume based trading system. The results show how the proposed Fuzzy Logic Momentum Analysis System outperforms the standard volume based systems that are widely used in the financial industry.

31.1 Introduction

The modelling of financial systems continues to hold great interest not only for researchers but also for investors and policymakers. Many of the characteristics of these systems, however, cannot be adequately captured by traditional financial modelling approaches. Financial systems are complex, nonlinear, dynamically changing systems in which it is often difficult to identify interdependent variables and their values.

In particular, the problem of optimal order execution has been a main concern for financial trading and brokerage firms for decades [1]. The idea of executing a

A. Kablan (✉) · W. L. Ng
Centre for Computational Finance and Economic Agents (CCFEA), University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK
e-mail: akabla@essex.ac.uk

W. L. Ng
e-mail: wlng@essex.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_31,
© Springer Science+Business Media B.V. 2011

client's order to buy or sell a pre-specified number of shares at a price better than all other competitors seems intriguing. However, this involves the implementation of a system that considers the whole price formation process from a different point of view. Financial brokers profit from executing clients' orders of buying and selling of certain amounts of shares at the best possible price. Many mathematical and algorithmic systems have been developed for this task [2], yet they seem not to be able to overcome a standard volume based system.

Most systems use well-documented technical indicators from financial theory for their observations. For example, [3] used three technical indicators in their stock trading system: the rate of change, the stochastic momentum indicator and a support-resistance indicator that is based on the 30-day price average. A convergence module then maps these indices as well as the closing price to a set of inputs for the fuzzy system, thus providing a total of seven inputs. In some cases, such as the rate of change, one indicator maps to a single input. However, it is also possible to map one indicator to multiple inputs. Four levels of quantification for each input value are used: small, medium, big and large. In this case, Mamdani's form of fuzzy rules [4] can be used to combine these inputs and produce a single output variable with a value between 0 and 100. Low values indicate a strong sell, high values a strong buy. The system is evaluated using 3 years of historical stock price data from four companies with variable performance during one period and employing two different strategies (risk-based and performance-based). In each strategy, the system begins with an initial investment of $10,000 and assumes a constant transaction cost of $10. Similarly, tax implications are not taken into consideration. The resulting system output is shown to compare favourably with stock price movement, outperforming the S&P 500 in the same period.

The application presented in this study differs from the above, as it introduces a fuzzy logic-based system for the momentum analysis [5]. The system uses fuzzy reasoning to analyse the current market conditions according to which a certain equity's price is currently moving. This is then used as a trading application. First, the membership functions were decided by the expert-based method, but then later optimised using ANFIS [6], further improving the trading strategy and order execution results.

31.2 Fuzzy Logic Momentum Analysis System (Fulmas)

31.2.1 Fuzzy Inference

A fuzzy inference system is a rule-based fuzzy system that can be seen as an associative memory and is comprised of five components:

- A rule base which consists of the fuzzy if–then rules.
- A database which defines membership functions of the fuzzy sets used in the fuzzy rules.

- A decision-making unit which is the core unit and is also known as the inference engine.
- A fuzzification interface which transforms crisp inputs into degrees of matching linguistic values.
- A defuzzification interface which transforms fuzzy results into crisp output.

Many types of fuzzy inference systems have been proposed in literature [7]. However, in the implementation of an inference system, the most common is the Sugeno model, which makes use of if–then rules to produce an output for each rule. Rule outputs consist of the linear combination of the input variables as well as a constant term; the final output is the weighted average of each rule's output. The rule base in the Sugeno model has rules of the form:

$$\text{If } X \text{ is } A_1 \text{ and } Y \text{ is } B_1, \quad \text{then} \quad f_1 = p_1 X + q_1 Y + r_1. \tag{31.1}$$

$$\text{If } X \text{ is } A_2 \text{ and } Y \text{ is } B_2, \quad \text{then} \quad f_2 = p_2 X + q_2 Y + r_2. \tag{31.2}$$

X and Y are predefined membership functions, A_i and B_i are membership values, and p_i, q_i and r_i are the consequent parameters. When we calculate the equation of first-order Sugeno model [8], the degree of membership variable of x_1 in membership function of A_i are multiplied by the degree of membership variable of x_2 and in membership function B_i, and the product is weight W_i. Finally, the weighted average of f_1 and f_2 is deemed the final output Z, which is calculated as

$$Z = \frac{W_1 \cdot f_1 + W_2 \cdot f_2}{W_1 + W_2}. \tag{31.3}$$

In the case of designing a fuzzy system for financial modelling, one should opt to use a model similar to Mamdani and Assilian [4], which is based on linguistic variables and linguistic output. Basically, fuzzy logic provides a reasoning-like mechanism that can be used for decision making. Combined with a neural network architecture, the resulting system is called a neuro-fuzzy system. Such systems are used for optimisation since they combine the reasoning mechanism that fuzzy logic offers together with the pattern recognition capabilities of neural networks, which will be discussed in the following.

31.2.2 Adaptive Neuro Fuzzy Inference System (ANFIS)

The ANFIS is an adaptive network of nodes and directional links with associated learning rules [6]. The approach learns the rules and membership functions from the data [8]. It is called adaptive because some or all of the nodes have parameters that affect the output of the node. These networks identify and learn relationships between inputs and outputs, and have high learning capability and membership function definition properties. Although adaptive networks cover a

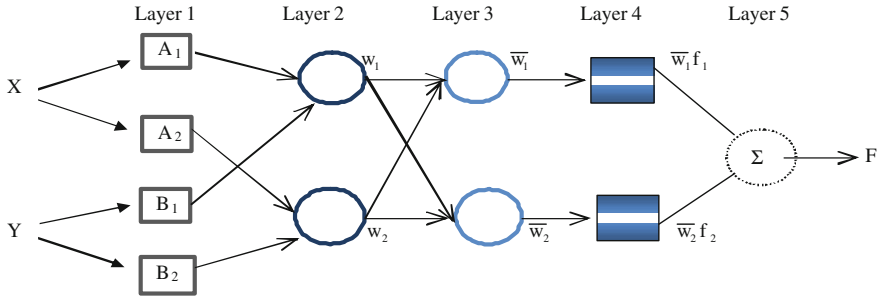


Fig. 31.1 ANFIS architecture for a two rule Sugeno system

number of different approaches, for our purposes, we will conduct a detailed investigation of the method proposed by Jang et al. [9] with the architecture shown in Fig. 31.1.

The circular nodes have a fixed input–output relation, whereas the square nodes have parameters to be learnt. Typical fuzzy rules are defined as a conditional statement in the form:

$$\text{If } X \text{ is } A_1, \quad \text{then } Y \text{ is } B_1 \tag{31.4}$$

$$\text{If } X \text{ is } A_2, \quad \text{then } Y \text{ is } B_2 \tag{31.5}$$

However, in ANFIS we use the 1st-order Takagi–Sugeno system [8] shown in Eq. 31.1 and 31.2. ANFIS can also be used to design forecasting systems [10]. We briefly discuss the five layers in the following:

1. The output of each node in Layer 1 is:

$$O_{1,i} = \mu_{A_i}(x) \quad \text{for } i = 1, 2$$

$$O_{1,i} = \mu_{B_{i-2}}(x) \quad \text{for } i = 3, 4$$

Hence, $O_{1,i}(x)$ is essentially the membership grade for x and y. Although the membership functions could be very flexible, experimental results lead to the conclusion that for the task of financial data training, the bell-shaped membership function is most appropriate (see, e.g., Abonyi et al. [11]). We calculate

$$\mu_A(x) = \frac{1}{1 + \left|\frac{x - c_i}{a_i}\right|^{2bi}}, \tag{31.6}$$

where a_i,b_i,c_i are parameters to be learnt. These are the premise parameters.

2. In Layer 2, every node is fixed. This is where the t-norm is used to "AND" the membership grades, for example, the product:

$$O_{2,i} = W_i = \mu_{A_i}(x)\mu_{B_i}(y), \quad i = 1, 2. \tag{31.7}$$

3. Layer 3 contains fixed nodes that calculate the ratio of the firing strengths of the rules:

$$O_{3,i} = \overline{W_i} = \frac{W_i}{W_1 + W_2}. \tag{31.8}$$

4. The nodes in Layer 4 are adaptive and perform the consequent of the rules:

$$O_{4,i} = \overline{W_i} f_i = \overline{W_i}(p_i x + q_i y + r_i). \tag{31.9}$$

   The parameters (p_i, q_i, r_i) in this layer are to be determined and are referred to as the consequent parameters.
5. In Layer 5, a single node computes the overall output:

$$O_{5,i} = \sum_i \overline{W_i} f_i = \frac{\sum_i W_i f_i}{\sum_i W_i} \tag{31.10}$$

This is how the input vector is typically fed through the network layer by layer. We then consider how the ANFIS learns the premise and consequent parameters for the membership functions and the rules in order to optimise these in the Fuzzy Logic Momentum Analysis System to produce a further improved system with a higher performance.

31.2.3 Fulmas for Trading

Creating a fuzzy inference system to detect momentum is a complex task. The identification of various market conditions has been a topic subject to various theories and suggestions [12]. In the following, the proposed fuzzy inference system categorises the market conditions into seven categories based on price movement, using the current volume to determine the participation rates (PR) of the trading system each time. The participation rate is the amount of volume that will be traded at each instance.

The first step in designing the Fuzzy Logic Momentum Analysis System involves defining the "market conditions" that the fuzzy system has to identify. The following seven market conditions are used to cover all possible movements of the price series:

- Rallying
- Strong up
- Slightly up
- Average
- Slightly down
- Strong down
- Crashing

These conditions are considered as linguistic values for the fuzzy logic system, and they will be used to determine the current state of the price formation and its momentum. As momentum builds, the system considers the previous x amount of ticks and performs an inference procedure by adding all of the movements of the current price to the previous price in order to determine whether the general trend has been up or down after x points.

Let P_i denote the current price and P_{i-1} the previous price; k_i is a fluctuating counter that goes up or down according to the movement of the price. Whenever the price goes up, it adds 1, and when the price goes down, it subtracts 1. Hence, this can be used to identify market conditions price movements, where if the market is moving strongly upwards, it will be detected by having more +1 than −1 or 0. This can be modelled as

$$\text{Momentum}(x) = \sum_{i=1}^{x} k_i \tag{31.11}$$

where x is the number of ticks where we want to detect the momentum. For example, if we want to detect the momentum of the last 100 ticks, we count all up and down movements and then feed the resulting number to the fuzzy system, whose output would lie somewhere in the membership functions. The choice of triangular membership functions was made after using the expert based method, where it was suggested that triangular membership functions should be used due to their mathematical simplicity. Triangular shapes require three parameters and are made of a closed interval and a kernel comprised of a singleton. This simplifies the choice of placing the membership functions. The expert merely has to choose the central value and the curve slope on either side.

The same procedure is applied for calculating the linguistic variable "volatility", where the linguistic values are:

- Very high
- High
- Medium
- Low
- Very low

The fuzzy logic system considers both market momentum and volatility. It generates the rules and then takes a decision based on the amount of market participation. This is illustrated in Fig. 31.2.

31.3 Empirical Analysis

Experiments in this study have been carried out on high-frequency tick data obtained from ICAP plc of both Vodafone Group plc (VOD) and Nokia Corporation (NOK). A very important characteristic of this type of data is that it is

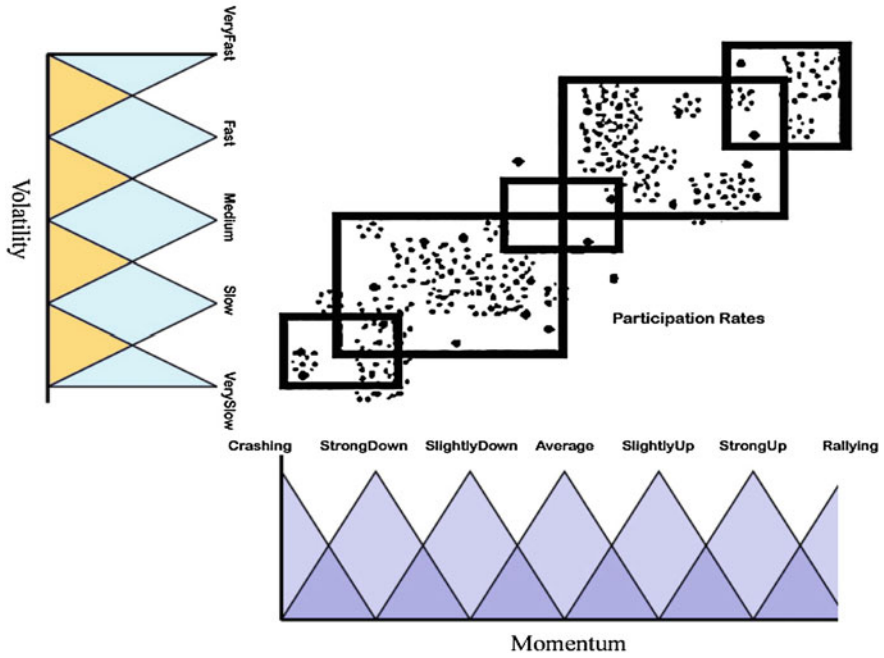


Fig. 31.2 Extracting fuzzy rules from both volatility and momentum

irregularly spaced in time, which means that the price observations (ticks) are taken in real-time (as they arrive). The application is designed for an interdealer broker,[1] which means that they have the ability to create orders with any amount of volume.

For both stocks, 2 months of high-frequency tick data between 2 January 2009 and 27 February 2009 has been obtained, simulations are terminated whenever 1 million shares have been bought or sold. The fuzzy logic system receives the first batch of data and performs all of the buy or sell actions on it. The same procedure is repeated using the standard volume-based system. Finally, the performance of both systems is compared.

It must be mentioned that 2 months of high-frequency tick data is a significantly large amount of data; considering every iteration, the system analyses the momentum of the past 100 ticks (Fig. 31.3).

[1] An interdealer broker is a member of a major stock exchange who is permitted to deal with market makers, rather than the public, and can sometimes act as a market maker.

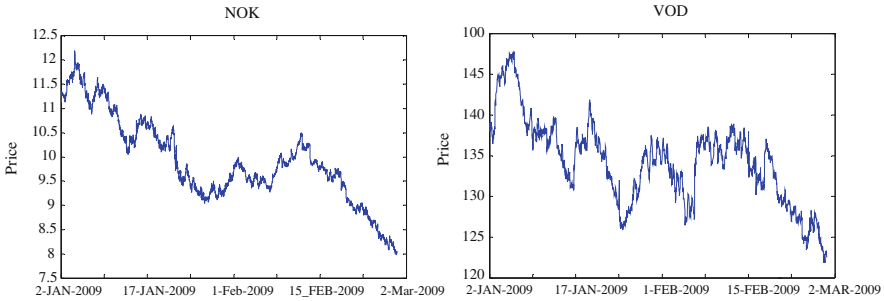


Fig. 31.3 Time series data of NOK and VOD prices

31.3.1 Standard Volume System (SVS)

A standard brokerage and trading mechanism for executing large orders is a simple volume-based system that parses the volume being traded whenever a certain number of shares (a threshold) have been traded; the system will buy or sell (depending on the order) a certain percentage. If there is an order to trade one million shares of a certain stock, the threshold could be, for example, 10,000 shares. Whenever 10,000 shares have been traded and if the participation rate PR is set to 25%, the system will buy or sell 25% of the average volume. If the accumulated sum of the volume exceeds the predefined threshold, then the amount of shares traded is equal to the PR multiplied by the current volume:

$$\text{Total SVS Cost} = \sum_{i=1}^{n} \text{price}_i * (\text{amount of shares}_i)$$

where n is the number of operations required to reach the target order (for example, 1 million shares). The above system has proven to be efficient and is being adopted by many financial brokerage and execution institutions [13].

31.3.2 Benchmark Performance Measures

Although many systems have used many different approaches such quantum modelling to determine the various participation rates (PR), they usually fail to outperform the standard volume system in the long term. The aim of this study is to prove that FULMAS outperforms this type of system in the long run, this is assessed using order execution costs for buy and sell orders. In particular, FULMAS will be applied to determine the PR in the market according to the current momentum. For example, for a buy order, it is preferable to increase the PR (number of shares bought at that time) when the price is low and to decrease the participation when the price is high. The idea here is to use the momentum analysis system to identify in what market condition we are currently residing in.

Table 31.1 Participation rates for buy side and the sell side of FULMAS

	Buying participation rates (%)	Selling participation rates (%)
Rallying	10	40
Strong up	15	35
Slightly up	20	30
Average	25	25
Slightly down	30	20
Strong down	35	15
Crashing	40	10

This will enable us to vary the PR, providing a trading advantage, since the system can trade aggressively when the condition is at an extreme. It would also minimise its trading when the condition is at another extreme. In other words, if we are selling 1 million shares, the system will make a trade whenever the threshold of volume has been exceeded. However, if the current market condition indicates that the price is very high or rallying, then we know that this is a suitable time to sell a lot of shares, for example, 40% of the current volume. The same concept applies when the momentum indicates that the price is strong down, which means that the system should sell a lower volume at this low price, for example, 15%. The reverse mechanism applies for buying shares. When the market is crashing, this is a good indicator that we should buy a large volume (40%), and when the price is at an average point, it would behave like the SVS system, i.e., buying 25% of the volume. This is shown in Table 31.1. The same procedure is applied to volatility and then combined with volume to produce the fuzzy rules.

When implementing SVS and FULMAS, the benchmark at which both systems will be compared against each other will be the outperformance of FULMAS on the SVS, expressed in basis points (one hundredth of 1%). To calculate the improvement (imp) for the buy and sell sides, the following formulas are used:

$$\text{impBuy} = \left(1 - \frac{\text{FULMAS price}}{\text{SVS price}}\right) * 10^4 \text{ bps}$$

$$\text{impSell} = \left(\frac{\text{FULMAS price}}{\text{SVS price}} - 1\right) * 10^4 \text{ bps}$$

where FULMAS price is the total cost of buying x amount of shares using FULMAS, and SVS price is the total cost of buying the same number of shares using the traditional SVS.

31.3.3 Results

The complimentary characteristics of neural networks and fuzzy inference systems have been recognised and the methodologies have been combined to create neuro-

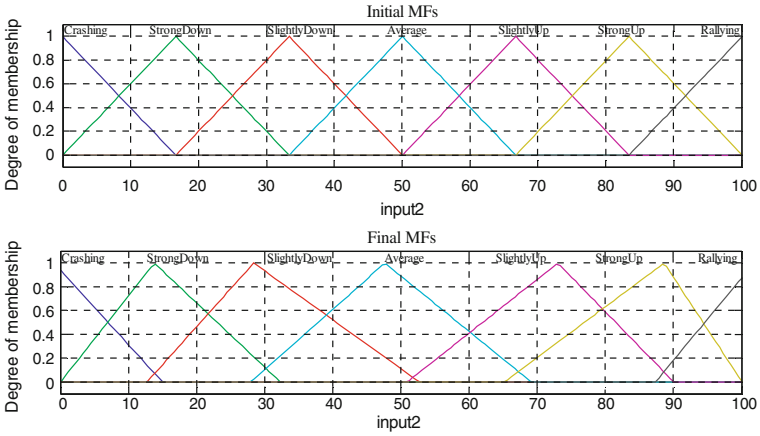


Fig. 31.4 Triangular membership functions optimised using ANFIS

fuzzy techniques. Indeed, earlier work by Wong and Wang [14] described an artificial neural network with processing elements that could handle fuzzy logic and probabilistic information, although the preliminary results were less than satisfactory. In this study, ANFIS is used to optimise the membership functions in FULMAS. This is performed by feeding the ANFIS system both the training data, the desired output, and tuning the ANFIS in order to reach the target result by modifying the membership functions (see Figs. 31.4 and 31.5). In other words, at each instance, ANFIS is fed the results currently obtained from the fuzzy system together with a set of target prices or data. This target price will be an optimal price that is far better than the current one (a cheaper price if on buy mode or a higher price if in sell mode). The system runs and modifies the membership

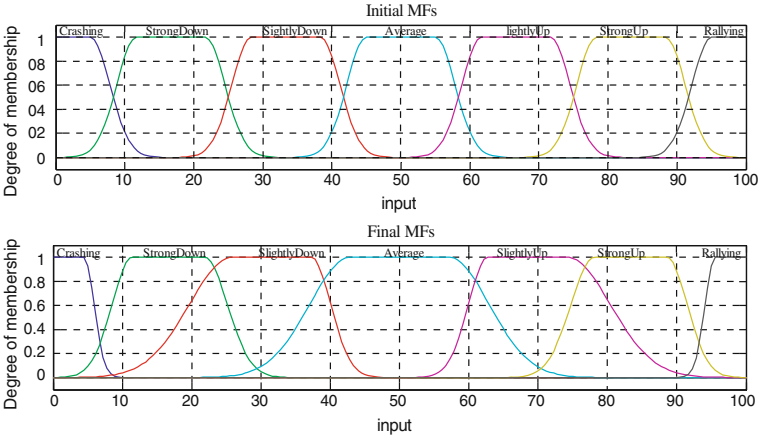


Fig. 31.5 Bell-shaped membership functions optimised using ANFIS

Table 31.2 Analysis of results of buying and selling 1 million shares of NOK and VOD with the descriptive statistics of the improvement indicators (in bps per trade)

	Mean	Median	Std dev	Skewness	Kurtosis
Initial results					
Buying NOK	2.98	4.63	12.39	−0.05	2.56
Buying VOD	12.48	1.58	36.25	1.74	4.86
Selling NOK	1.68	2.92	8.79	−1.43	6.25
Selling VOD	2.73	2.46	27.71	0.70	8.84
Optimised results					
Buying NOK	6.94	6.57	12.99	0.15	2.53
Buying VOD	14.48	4.33	2.95	−0.74	3.28
Selling NOK	9.36	5.79	9.18	−0.52	2.61
Selling VOD	7.71	6.91	28.23	0.86	9.38

functions in each epoch in order to get as close to the optimal price as possible. Comparing the results of both optimised membership functions, an improvement in the original system was discovered. The optimised triangular membership functions have also outperformed the optimised bell-shaped membership functions; this confirms the experts' opinion mentioned above concerning the choice of the triangular membership functions.

Table 31.2 displays the improvement of FULMAS against SVS, showing the descriptive statistics of the improvement rate of buying or selling one million. This improvement rate can be either positive, when FULMAS has outperformed SVS, or negative, when FULMAS was outperformed by SVS. In particular, we see a much higher outperformance than in the previous system, which confirms that the use of ANFIS to optimise the membership functions has increased the performance of the system on both the buy and sell sides. For example, Table 31.2 shows that on the buying side, the system, on average, outperforms the standard system by more than six basis points. On an industrial scale, this means a large amount of savings for financial institutions that employ such systems to vary the participation rates. Other descriptive statistics such as the standard deviation, skewness and kurtosis are also included. These imply that the outperformance of FULMAS over SVS is actually considerable given the higher values of the median. Also, the skewness is closer to zero, and the kurtosis has decreased in most cases, both implying a higher accuracy of the improved system.

31.4 Summary and Discussion

It is well known that a main inadequacy of economic theory is that it postulates exact functional relationships between variables. In empirical financial analysis, data points rarely lie exactly on straight lines or smooth functions. Ormerod [15] suggests that attempting to accommodate these nonlinear phenomena will introduce an unacceptable level of instability in models. As a result of this

intractability, researchers and investors are turning to artificial intelligence techniques to better inform their models, creating decision support systems that can help a human user better understand complex financial systems such as stock markets. Artificial intelligence systems in portfolio selection have been shown to have a performance edge over the human portfolio manager and recent research suggests that approaches incorporating artificial intelligence techniques are also likely to outperform classical financial models [16].

This study has introduced a system that utilises fuzzy logic in order to justify the current market condition that is produced by the accumulation of momentum. FULMAS is a fuzzy logic momentum analysis system that outperforms the traditional systems used in industry, which are often based on executing orders dependent on the weighted average of the current volume. Results of the implemented system have been displayed and compared against the traditional system. The system proves that, on average, it increases profitability on orders on both the buy and sell sides. FULMAS has been improved further by using ANFIS as an optimisation tool and the new results have shown a significant improvement over both the original FULMAS system and the SVS system.

Acknowledgments The authors would like to thank Mr. Phil Hodey, the head of portfolio management and electronic trading at ICAP plc for providing the tick data used in the simulations of the system and for his invaluable support and guidance.

References

1. Ellul A, Holden CW, Jain P, Jennings RH (2007) Order dynamics: recent evidence from the NYSE. J Empirical Finance 14(5):636–661
2. Chu HH, Chen TL, Cheng CH, Huang CC (2009) Fuzzy dual-factor time-series for stock index forecasting. Expert Syst Appl 36(1):165–171
3. Dourra H, Siy P (2002) Investment using technical analysis and fuzzy logic. Fuzzy Sets Syst 127(2):221–240
4. Mamdani E, Assilian S (1975) An experiment in linguistic synthesis with a fuzzy logic controller. Int J Man Mach Stud 7(1):1–13
5. Kablan A, Ng WL (2010) High frequency trading using fuzzy momentum analysis. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, vol I, 30 June–2 July, London, UK, pp 352–357
6. Jang JR (1993) ANFIS: adaptive network-based fuzzy inference system. IEEE Trans Syst Man Cybern 23(3):665–685
7. Dimitrov V, Korotkich V (2002) Fuzzy logic: a framework for the new millennium, studies in fuzziness and soft computing, vol 81. Springer, New York
8. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its application to modeling and control. IEEE Trans Syst Man Cybern 15(1):116–132
9. Jang JR, Sun CT, Mizutani E (1997) Neuro-fuzzy and soft computing. Prentice Hall, Upper Saddle River
10. Atsalakis GS, Valavanis KP (2009) Forecasting stock market short-term trends using a neuro-fuzzy based methodology. Expert Syst Appl 36(7):10696–10707
11. Abonyi J, Babuska R, Szeifert F (2001) Fuzzy modeling with multivariate membership functions: gray box identification and control design. IEEE Trans Syst Man Cybern B 31(5):755–767

12. Griffin J (2007) Do investors trade more when stocks have performed well? Evidence from 46 countries. Rev Financ Stud 20(3):905–951
13. Goldstein MA, Irvine P, Kandel E, Wiener Z (2009) Brokerage commissions and institutional trading patterns. Rev Financ Stud 22(12):5175–5212
14. Wong FS, Wang PZ (1990) A stock selection strategy using fuzzy neural networks. Neurocomputing 2(5):233–242
15. Ormerod P (2000) Butterfly economics: a new general theory of social and economic behaviour. Pantheon, New York
16. Brabazon A, O'Neill M, Maringer D (2010) Natural computing in computational finance, vol 3. Springer, Berlin

Chapter 32
The Determination of a Dynamic Cut-Off Grade for the Mining Industry

P. V. Johnson, G. W. Evatt, P. W. Duck and S. D. Howell

Abstract Prior to extraction from a mine, a pit is usually divided up into 3-D 'blocks' which contain varying levels of estimated ore-grades. From these, the order (or 'pathway') of extraction is decided, and this order of extraction can remain unchanged for several years. However, because commodity prices are uncertain, once each block is extracted from the mine, the company must decide in real-time whether the ore grade is high enough to warrant processing the block further in readiness for sale, or simply to waste the block. This paper first shows how the optimal cut-off ore grade—the level below which a block should be wasted—is not simply a function of the current commodity price and the ore grade, but also a function of the ore-grades of subsequent blocks, the costs of processing, and the bounds on the rates of processing and extraction. Secondly, the paper applies a stochastic price uncertainty, and shows how to derive an efficient mathematical algorithm to calculate and operate a dynamic optimal cut-off grade criterion throughout the extraction process, allowing the mine operator to respond to future market movements. The model is applied to a real mine composed of some 60,000 blocks, and shows that an extra 10% of value can be created by implementing such an optimal regime.

P. V. Johnson (✉) · G. W. Evatt · P. W. Duck
School of Mathematics, University of Manchester, Manchester, UK
e-mail: paul.johnson-2@manchester.ac.uk

G. W. Evatt
e-mail: geoffrey.evatt@manchester.ac.uk

P. W. Duck
e-mail: peter.duck@manchester.ac.uk

S. D. Howell
Manchester Business School, University of Manchester, Manchester, UK
e-mail: syd.howell@mbs.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_32,
© Springer Science+Business Media B.V. 2011

32.1 Introduction

Mineral mining is a complex engineering operation, which can last for several decades. As such, significant consideration must be given to the planning and design of the operation, so that numerous engineering constraints can be met, whilst making sure the operation is economically viable. To compound the difficulty of the task, the planning and scheduling of extraction from a mine is made in the presence of uncertainties, such as the future commodity price and estimated ore-grade. These uncertainties can fluctuate on a daily basis, highlighting the different timescales upon which the mining company must base decisions: the shorter time scales governed by commodity price and realised ore-grade, and the longer time-scales governed by (amongst other things) extraction rates and processing capacities. The focus of this paper is upon one of these short time-scale decisions: whether to process the extracted material, or to waste it. The level of ore-grade which separates this decision is known as the 'cut-off grade' [7].

Prior to extraction, the planning of the extraction schedule begins with deciding an appropriate pathway (or order) through the mine. Whilst it is possible to alter the order of extraction at various points during extraction, it is generally not a particularly flexible decision, as changing an order can require moving extraction machinery, processing units, the cancellation of contracts and large overhead costs. As such, it is reasonable to assume that the pathway through the mine is fixed, but it is how one progresses, and operates, along that pathway that is variable. At this planning stage, the mine is graphically divided up into 3-D blocks, each containing its own estimated quantity of ore. The estimated ore-grade carries with it an associated uncertainty, which can have an effect upon the valuation of a mining operation [6]. However, it is the expected (estimated) ore grade level which dominates the planning of the actual pathway through the mine, as this is the best-guess in deciding the order in which the resource should be extracted. The extraction pathway is most commonly decided using software such as the Gemcom-Whittle package [15], which allows companies to construct feasible pit shapes that satisfies slope constraints on the angle of the pit, transportation needs and work-force limitations. As previously mentioned, this algorithm may be used several times throughout a mine's life, so as to ensure the mine plan is consistent with market conditions, however on a day-to-day basis the mine must take more detailed scheduling decisions in real-time.

The key real-time decision is whether or not to process the latest extracted block (e.g. by milling or electrolysis) in readiness for sale, where the block's intrinsic value varies with its ore grade and with the underlying commodity price. We define a 'cost-effective' block as one whose ore grade is high enough to pay the cash costs of processing, at the current price. However the cut-off ore grade—above which a block should be processed—need not be set as low as the grade above which the block will be cost-effective to process. Disparity between the rate of extraction and the maximum processing capacity means that there can be an opportunity cost to processing all cost-effective material, since the small

short-term gain of processing a low grade block could be surpassed by bringing forward the processing of more valuable blocks instead. The optimal wasting of potentially cost-effective material is the focus of this paper.

To highlight the above point, let us consider a trivial case where the mine has a stock of 3 blocks awaiting processing, extracted in order, A, B and C, whose current market values after processing costs are $V_A = \$1$, $V_B = \$50$, and $V_C = \$1,000$. Whilst, classically, analysis has often been indifferent to the order of processing, with enough discounting applied one can see that by an optimal cut-off criterion, it would be best to simply waste A and get on with processing B and C. This is because the value gained in processing A is less than the time value of money lost in waiting to process B and C at a later date. This lack of consideration of the discount rate has been highlighted before as a drawback in current mine planning [14] but, as yet, little progress has been made with it.

Another consequence of an optimal cut-off grade decision is having to increase the rate of extraction of poor quality ores to keep the processing plant loaded. This is because a processing unit will typically operate at a fixed capacity, and closing (or restarting) it is a costly and undesirable operation. As such, a maximum (and minimum) possible extraction rate must be known. This clearly illustrates the link between extraction rate and the optimal cut-off grade. With this maximum possible extraction rate, one knows precisely which blocks can possibly be extracted within each period in time, and thus the decision as to which block to process next can be decided.

There have been several other approaches to mine valuation and the corresponding extraction regime. Typically these have relied upon simulation methods to capture the uncertainty of price and ore-grade [8, 9, 12]. These types of method can be extremely time consuming, with computing times of several hours [3], and can often lead to sub-optimal and incomplete results. Using these simulation techniques, optimal cut-off grades were investigated by Menabde et al. [10], although little insight into the core dynamics, performance or robustness was obtained. A similar approach is the use of genetic algorithms—a general technique commonly used by computer scientists—which are capable of calculating mine schedules whilst adhering to specified constraints upon their design [11]. Whilst the work of Myburgh and Deb [11] was suitable in calculating feasible paths, the criteria by which this particular study operated was, again, not given, and the computing time was also of the order of hours.

To make a step-change away from these methods, partial differential equations (PDEs) can be implemented to capture the full mine optimisation process, which builds on work by Brennan and Schwartz [2] and Chen and Forsyth [4]. The inclusion of stochastic ore-grade uncertainty, via PDEs has also been tested by Evatt et al. [6], which enabled mine valuations to be produced in under 10 s and showed that the effect on mine value of stochastic ore-grade variation is much less than the effect of stochastic price. Whilst the mathematics and numerics of this PDE approach are relatively complex at the outset, once solved, they produce highly accurate results in short times—complete with model input sensitivities. This paper extends the use of PDEs, adding a model for tactical processing

decisions under foreseeable variations in ore grade and unforeseeable fluctuations in price. This shows that when processing capacity is constrained, the ability to maximise the value of processing by varying the cut-off ore grade can add significantly to mine value when optimally applied. By solving rapidly under a range of processing constraints, the scale of the processing plant can itself be optimised.

In Sect. 32.2 we demonstrate the underlying concepts determining the optimal cut-off decision rule, and in Sect. 32.3 we apply a price uncertainty to the model and use a contingent claims approach to derive the governing equation. We then apply the model to a mine composed of some 60,000 blocks in Sect. 32.4, to show how much extra value the running of an optimal cut-off grade regime can add to a valuation. We draw together our concluding remarks in Sect. 32.5.

32.2 Cut-Off Grade Optimisation

The selection of the cut-off grade criteria reduces to whether a cost-effective block should be processed or not. This is because there is the possibility a more valuable block could be brought forward in time to be processed, which otherwise would loose more time-value of money than the value gained from processing the first block. To highlight this point let us consider the order of extracted of blocks from a mine, which we (hypothetically) place in a chronologically ordered row. As we operate the processing unit of the mine, we must pass along this row and decide which blocks to process and which blocks to waste. In reality, although we know the (estimated) ore-grades of the blocks in advance, until we know for certain the market price at the time of processing we cannot know what cashflow it will generate. Yet even if we assume a constant price, we can still show how dynamic cut-off grade decision making is still required and optimal.

Consider a highly simplified mine, as shown in Fig. 32.1, which is composed of just two blocks, Block1 and Block2, with ore grades G_1 and G_2, respectively. We allow the mine to have the capacity within the rate of extraction to immediately process either the first block, Block1, or its successor, Block2. As such, the comparison is between the value of processing both blocks in order, given by V_{12}, or the value of only processing Block2, V_2. With a constant price, S, we can write down the net present value of these two (already extracted) blocks, where we shall process both,

$$V_{12} = (SG_1 - \epsilon_P) + (SG_2 - \epsilon_P)e^{-r\delta t}. \tag{32.1}$$

Here δt is time it takes to process each block, ϵ_P is the cost of processing each block and the discount rate is r. This value must be compared to the decision to waste the first block and process only the second block, which would have value,

$$V_2 = (SG_2 - \epsilon_P). \tag{32.2}$$

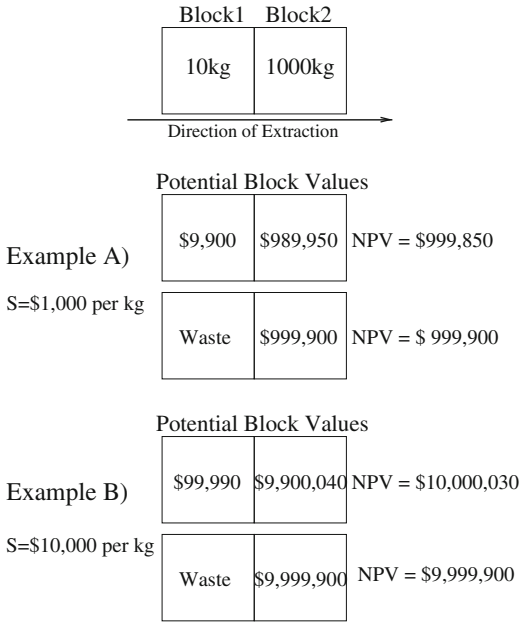


Fig. 32.1 Two examples of how price may effect the order in which blocks are processed so as to maximise a mines NPV. Example A is made with a low commodity price, $S = \$1,000\,\mathrm{kg}^{-1}$, and Example B is made with a high commodity price, $S = \$10,000\,\mathrm{kg}^{-1}$

This comparison between V_{12} and V_2 is one the algorithm must continually make. To demonstrate how the selection depends upon the underlying price, Fig. 32.1 shows the choices available for two different commodity prices, one high ($S = \$10,000\,\mathrm{kg}^{-1}$) and one low ($S = \$1,000\,\mathrm{kg}^{-1}$). These are made with prescribed parameter values

$$r = 10\%, \quad \epsilon_P = \$100\,\mathrm{block}^{-1}, \quad \delta t = 0.1\,\mathrm{year}. \tag{32.3}$$

As can be seen, in the low-price case, Example A, it is best to process only the second block. However, in the high commodity price case, namely Example B, it is best to process both blocks. This simple example demonstrates (albeit with rather exaggerated parameter values) how the selection needs to be actively taken, and how different values of the underlying price, and discount rate, will affect the optimal cut-off decision. Another consequence of this optimal decision taking is that the mine will be exhausted earlier than might have been previously thought, since we wasted the first block and only processed the second, hence a mine owner could agree a shorter lease on this particular mine.

32.3 Model Construction

To create the framework for determining an optimal dynamic cut-off grade, we can make use of two distinct methods for arriving at the core equation describing the valuation, V. The first method follows a *contingents claims* approach, in which the

uncertainty arising from the underlying price is removed by hedging away the risk via short-selling suitable quantities of the underlying resource. The second method follows the *Feynman–Kac* probabilistic method, as described in relation to the mining industry by Evatt et al. [5], which is the chosen method for deriving a valuation when hedging is not undertaken. This second method is also permissible when hedging does take place but a slight adjustment to the price process is required, and explained within this latter paper. Because Evatt et al. [5] already covers the derivation of the mine valuation, in the present paper we explain how the contingent claims approach can be used.

We first prescribe three state-space variables; these are the price per unit of the underlying resource in the ore S, the remaining amount of ore within the mine Q and time t. We next need to define the underlying price uncertainty process, which we assume to follow a geometric Brownian motion,

$$dS = \mu S\, dt + \sigma_s S\, dX_s, \tag{32.4}$$

where μ is the drift, σ_s the volatility of S and the random variable dX_s, is a standard Wiener process. We use this price process without loss of generality, since other price processes (such as mean-reverting Brownian motion) can easily be implemented by the techniques described here.

Using the contingent claims approach (see [16]) and the above notation, we may apply Ito's lemma to write an incremental diffusive change in V as

$$dV = \sigma_s \frac{\partial V}{\partial S} dX_s + \frac{\partial V}{\partial Q} dQ + \left(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma_s^2 \frac{\partial^2 V}{\partial S^2} + \mu \frac{\partial V}{\partial S} \right) dt, \tag{32.5}$$

where we have taken powers of $(dt)^2$ and $(dQ)^2$ to be negligible. We are able to remove the dQ term via the relationship between Q and t by specifying the rate of extraction, q_e, namely,

$$dQ = -q_e\, dt, \tag{32.6}$$

where q_e can be a function of all three variables, if required. This extraction rate is the function we wish to determine in our optimal cut-off regime, as it governs both how we progress through the mine and, as a consequence, which blocks we choose to waste. The rate of extraction will obviously have limitations on its operating capacity, $q_e \in [0, q_{max}]$, which itself could be a function of time. The rate of extraction is closely linked to the rate of processing, which should be kept at a fixed constant, q_p. Hence q_{max} must be big enough for the processing unit to always operate at its constant capacity, q_p, i.e. there must always be enough cost-effective ore-bearing material being extracted from the mine so as to meet the processing capacity. Optimal variation in the extraction rate has already been shown to produce improved valuations [7], although this was achieved without considering processing limitations or grade variation.

With this relationship, (32.6), Eq. 32.5 can be transformed into

$$dV = \sigma_1 \frac{\partial V}{\partial S} dX_s + \left(\frac{\partial V}{\partial t} - q_e \frac{\partial V}{\partial Q} + \frac{1}{2} \sigma_s^2 \frac{\partial^2 V}{\partial S^2} + \mu \frac{\partial V}{\partial S} \right) dt. \tag{32.7}$$

To follow the conventional approach in creating and valuing risk-free portfolios we construct a portfolio, Π, in which we are instantaneously long in (owning) the mine and short in (owing) γ_s amounts of commodity contracts. This defines $\Pi = V - \gamma_s S$, such that,

$$d\Pi = dV - \gamma_s \, dS. \tag{32.8}$$

This portfolio is designed to contain enough freedom in γ_s to be able to continually hedge away the uncertainty of dX_s, which is the standard approach in creating risk-free portfolios [1, 13]. It also implies that within a small time increment, dt, the value of Π will increase by the risk-free rate of interest, minus any economic value generated and paid out by the mine during the increment. This economic value is typically composed of two parts, the first, negative, being the cost to extract, $q_e \epsilon_M$, and the second, positive, the cash generated by selling the resource content of the ore processed, $q_p(SG - \epsilon_P)$. Here ϵ_M is the cost of extraction per ore tonne, ϵ_P is the processing cost per ore tonne, and G is the ore-grade (weight of commodity per ore tonne). The reason why the economic functions contain the factors q_e or q_p is that we wish to maximise value by varying q_e in real time, so as to maintain q_p at its fixed bound. In turning the discrete block model into a continuous function describing the ore grade, G, we have assumed that blocks are small enough that they can be approximated as infinitesimal increments of volume.

As discussed in Sect. 32.2, the decision whether to process or waste the next block must be optimised. Before or after optimisation the incremental change in portfolio value may be written as

$$d\Pi = r\Pi \, dt - \gamma_S \delta S \, dt - q_p(GS - \epsilon_P) \, dt - q_e \epsilon_M \, dt. \tag{32.9}$$

By setting the appropriate value of γ_s to be

$$\gamma_s = \frac{\partial V}{\partial S},$$

and substituting Eqs. (32.4), (32.7) and (32.8) into (32.9), we may write our mine valuation equation as

$$\begin{aligned} &\frac{1}{2} \sigma_s^2 S^2 \frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t} - q_e \frac{\partial V}{\partial Q} + (r - \delta) S \frac{\partial V}{\partial S} \\ &\quad - rV + q_p(GS - \epsilon_P) - q_e \epsilon_M = 0. \end{aligned} \tag{32.10}$$

This is of the same form as that derived by Brennan and Schwartz [2], except that they added taxation terms, but did not model processing constraints or variations of ore grade.

We next need to prescribe boundary conditions for (32.10). The boundary condition that no more profit is possible occurs either when the reserve is exhausted $Q = 0$, or when a lease to operate the mine has reached its expiry date $t = T$, hence:

$$V = 0 \quad \text{on } Q = 0 \quad \text{and/or} \quad t = T. \tag{32.11}$$

Since the extraction rate will have a physical upper bound, the extraction rate and cost will not vary with S when S is large. This permits a far field condition of the form

$$\frac{\partial V}{\partial S} \to A(Q, t) \quad \text{as } S \to \infty. \tag{32.12}$$

When the underlying resource price is zero we need only solve the reduced form of Eq. 32.10 with $S = 0$, which reduces to

$$V = e^{-rt} \int_0^T q_e \epsilon_M(z) e^{rz}\, dz. \tag{32.13}$$

This completes the determination of our core equation, and its boundary conditions. We can now define the optimising problem which we wish to solve: we must determine the optimal extraction rate, q_e^*, at every point in the state space which maximises the value V, which satisfies Eq. 32.10, with $q_e = q_e^*$, subject to the defined boundary conditions. Problems of this type may be solved numerically using finite-difference methods, in particular the semi-Lagrangian numerical technique (see [4] for further details). All results in this paper have been thoroughly tested for numerical convergence and stability.

We must now show how the optimal q^* and its corresponding cut-off grade is to be incorporated into the maximisation procedure.

32.4 Example Valuation

We now apply our optimal cut-off grade model to a real mine of some 60,000 blocks, whose block by block ore-grade and sequence of extraction were supplied by Gemcom Software International. This mine has an initial capital expenditure of some \$250m. We were also supplied with a fixed reference price S_{ref}, for us to compare valuations with. We ourselves assumed a maximum extraction rate of five times the processing rate, which is broadly realistic, and it restricts the mine to wasting no more than 80% of any section of cost-effective ore (if one can increase the extraction rate fivefold, then it is possible to waste four blocks and process the fifth). The other parameter values we were supplied are

$$r = 10\%\,\text{year}^{-1}, \quad \delta = 10\%\,\text{year}^{-1}, \quad \sigma_s = 30\%\,\text{year}^{-\frac{1}{2}},$$
$$S_{ref} = \$11,800\,\text{kg}^{-1}, \quad \epsilon_P = \$4\,\text{tonne}^{-1}, \quad \epsilon_e = \$1\,\text{tonne}^{-1},$$
$$Q_{max} = 305,000,000\,\text{tonnes}, \quad q_p = 20,000,000\,\text{tonnes year}^{-1}. \quad (32.14)$$

Whilst the ore-grade is quite volatile, it was shown in Evatt et al. [6] that a suitable average of the estimated grade quality could be used without any sizeable alteration in the valuation, as one would expect, since the same volume of ore is available sold whether one takes average values or not. Using this average, Fig. 32.2 shows the economic worth throughout extraction for each part of the mine, where we have assumed the price to remain at its prescribed reference price, $S_{ref}G - \epsilon_P$. This highlights how the grade varies through the extraction process, and it is with reference to this grade variation that we shall compare the regions where it is optimal to speed up extraction and consequently waste certain parts of the ore body.

32.4.1 Results

For the example mine, we first calculate and compare two different valuations made with and without the optimal cut-off criterion. Figure 32.3 shows two sets of valuations: the lower pair (straight lines; one dashed, one solid) shows the valuations made assuming a constant price ($\sigma_s = 0\%$), and the upper pair (curved lines; one dashed, one solid) shows the effect of including both price uncertainty ($\sigma_s = 30\%$) and the option to abandon the mine when the valuation becomes negative—which is a standard option to include in a reserve valuation [2]. In each pair of lines the lower, dotted lines show valuation without a cut-off regime, and the higher, solid lines show valuation with the optimal cut-off regime. The optimal cut-off regime increases the mine valuation by up to 10%, with increasing benefit

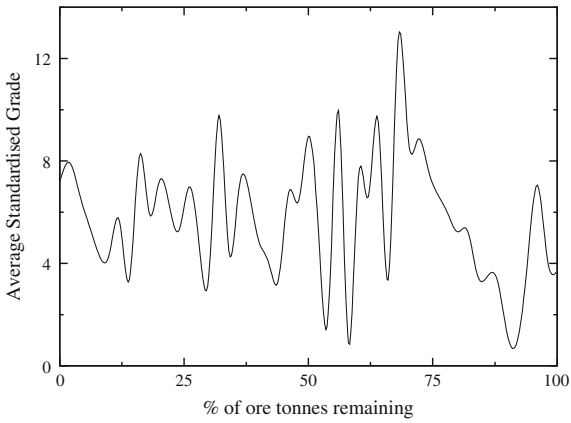


Fig. 32.2 Given a block ordering in the mine, the average standardised grade value is the cash value of ore (against reference price) minus processing costs per tonne of ore. This data was supplied by Gemcom Software International

at higher prices. This may seem surprising, but although the mine is always more profitable at higher prices, the opportunity cost of not allocating the finite processing capacity to the best available block does itself grow.

An obvious question which arises from this analysis is how do we decide which ore-grades we should waste, and what is the corresponding rate of extraction to achieve this? Given the mine operator will know at each point in time what the current underlying price is, they can look at the corresponding slice through the 3-D surface of the optimal cut-off grade, and see for which regions in t and Q they would waste ore and increase the rate of extraction. With this we can refer back to the corresponding grade of Fig. 32.2 and easily calculate what these grades actually are. For example, by looking at the closed regions of Fig. 32.4 we can see the optimal cut-off grades for two different commodity prices, $S = 100\%$ (top) and $S = 200\%$ (bottom) of the reference price. The points at where it is optimal to increase the rate of extraction is given by the segments where the closed regions (bounded by the thin line) intersect with the optimal extraction trajectory (bold line). In the two examples of Fig. 32.4, both appear to correspond to a standardised cut-off grade (Fig. 32.2) of around 2 units. The optimal rate of extraction is given by the gradient of the bold line, where the trajectory is calculated by integrating (32.6) for a given extraction regime. The difference between the dotted line (trajectory for the no cut-off situation), and the thick straight line of the optimal cut-off regime therefore gives an indication of the total amount of ore wasted.

Finally, Fig. 32.5 shows how the NPV depends upon the expected expiry time for extraction if one operates an optimal cut-off regime (solid line) or not (dotted line). If the mine chooses the optimal regime, the maximum NPV occurs just after 14 years, as opposed to the life of the mine being maximal at mine exhaustion at 15 years (as it is with no cut-off). This is a consequence of an optimal cut-off grade regime, in which the mine will occasionally increase its extraction rate from the (originally) planned level due to market fluctuations, thereby reaching the final pit shape in a shorter time.

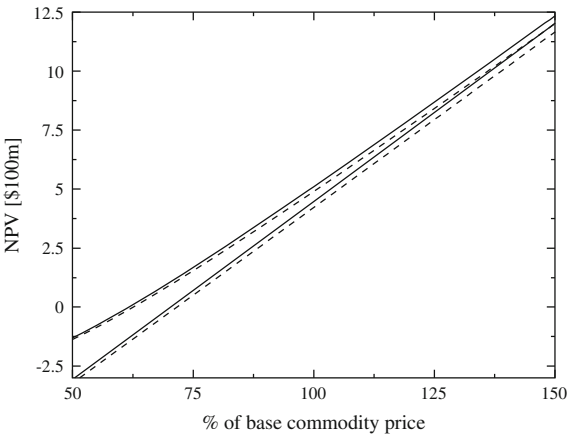


Fig. 32.3 NPV of the mine against percentage of reference price for two different sets of valuations. The two *lower lines* (*straight lines*; *one dashed*, *one solid*) are for a constant price while the two *upper lines* (*curved lines*; *one dashed*, *one solid*) include price volatility and the abandonment option. NPV for the optimal cut-off regime is shown by *solid lines*, and no cut-off by *dashed lines*

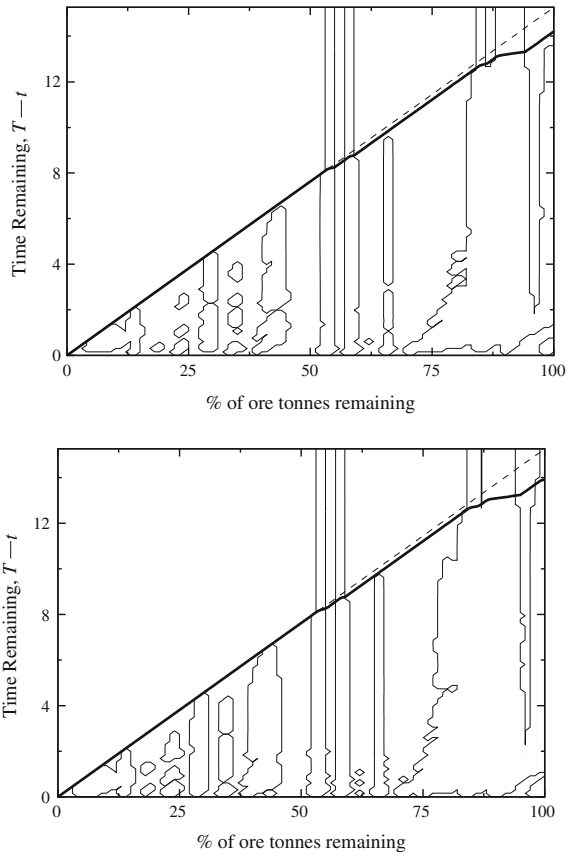


Fig. 32.4 Graphs showing the optimal cut-off regions for an extraction project for two different price levels, medium (*top*), and high (*bottom*). The closed regions contained within the *thin solid lines* show where ore is wasted and the extraction rate is increased. The *dashed line* represents the one realisation of a trajectory followed with no cut-off, while the *thick solid line* represents the realisation of the trajectory followed with optimal cut-off

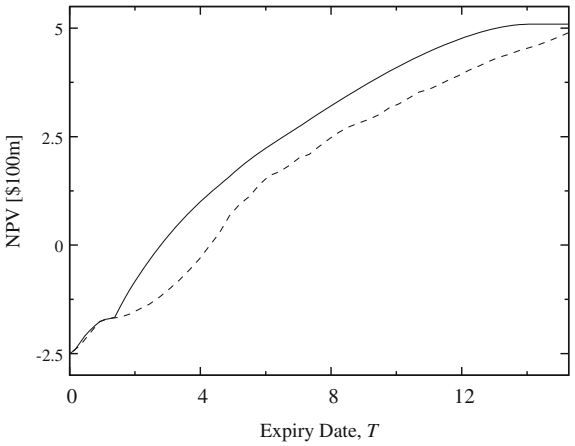


Fig. 32.5 The NPV of the mine against time remaining on the option on the mine given that 100% of the mine is present. The *solid line* is with optimal cut-off, *dashed* without

32.5 Conclusions

This paper has shown how to solve and optimise a (relatively) short time-scale mining problem, known as a dynamic cut-off grade, which is the continuous decision of whether to process extracted ore or not. This was achieved in the presence of price uncertainty. We have described how the partial differential equation model can be derived via two distinct methods, either by a contingent claims approach, when continuous hedging is present, or by the Feynman–Kac method. Using this model, we have shown how to determine and operate a optimal dynamic cut-off grade regime. As such, we have valued the 'option' to process or not to process under uncertainty, allowing the mine owner to react to future market conditions.

With our given example, the option adds around 10% to the expected NPV of an actual mine of 60,000 blocks. One natural extension of this work will be to allow for the cut-off grade to remain fixed for discrete periods of time, thus allowing mine operators to not have to continually alter their rate of extraction due to market changes.

References

1. Black F (1976) The pricing of commodity contracts. J Financial Econ 3:167–179
2. Brennan MJ, Schwartz ES (1985) Evaluating natural resource investments. J Business 58(2):135–157
3. Caccetta L, Hill SP (2003) An application of branch and cut to open pit mine scheduling. J Global Optim 27:349–365
4. Chen Z, Forsyth PA (2007) A semi-Lagrangian approach for natural gas storage valuation and optimal operation. SIAM J Sci Comput 30(1):339–368
5. Evatt GW, Johnson PV, Duck PW, Howell SD, Moriarty J (2010) The expected lifetime of an extraction project. In: Proceedings of the Royal Society A, Firstcite. doi:10.1098/rspa.2010.0247
6. Evatt GW, Johnson PV, Duck PW, Howell SD (2010) Mine valuations in the presence of a stochastic ore-grade. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, vol III, WCE 2010, 30 June–2 July, 2010, London, UK, pp 1811–1866
7. Johnson PV, Evatt GW, Duck PW, Howell SD (2010) The derivation and impact of an optimal cut-off grade regime upon mine valuation. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, WCE 2010, 30 June–2 July, 2010, London, UK, pp 358–364
8. Jewbali A, Dimitrakopoulos R (2009) Stochastic mine planning—example and value from integrating long- and short-term mine planning through simulated grade control. Orebody modelling and strategic mine planning, 2nd edn. The Australasian Institute of Mining and Metallurgy, Melbourne, pp 327–333
9. Martinez LA (2009) Designing, planning and evaluating a gold mine project under in-situ metal grade and metal price uncertainties. Orebody modelling and strategic mine planning, 2nd edn. The Australasian Institute of Mining and Metallurgy, Melbourne, pp 225–234
10. Menabde M, Foyland G, Stone P, Yeates GA (2004) Mining schedule optimisation for conditionally simulated orebodies. In: Proceedings of the international symposium on orebody modelling and strategic mine planning: uncertainty and risk management, pp 347–52

11. Myburgh C, Deb K (2010) Evolutionary algorithms in large-scale open pit mine scheduling. In: Proceedings of the 12th annual conference on genetic and evolutionary computation, pp 1155–1162
12. Ramazan S, Dimitrakopoulos R (2007) Stochastic optimisation of long-term production scheduling for open pit mines with a new integer programming formulation. Orebody modelling and strategic mine planning. The Australasian Institute of Mining and Metallurgy, Melbourne, pp 385–391
13. Schwartz ES (1997) The stochastic behavior of commodity prices: implications for valuation and hedging. J Finance LII(3):923–973
14. Tolwinski B, Underwood R (1996) A scheduling algorithm for open pit mines. IMA J Math Appl Bus Ind 7:247–270
15. Whittle D, Cahill J (2001) Who plans mines? In: Strategic mine planning conference, Perth, WA, pp 15–18
16. Wilmott P, Howison S, Dewynne J (1995) The mathematics of financial derivatives. Cambridge University Press, Cambridge

Chapter 33
Improved Prediction of Financial Market Cycles with Artificial Neural Network and Markov Regime Switching

David Liu and Lei Zhang

Abstract This paper provides an analysis of the Shanghai Stock Exchange Composite Index Movement Forecasting for the period 1999–2009 using two competing non-linear models, univariate Markov Regime Switching model and Artificial Neural Network Model (RBF). The experiment shows that RBF is a useful method for forecasting the regime duration of the Moving Trends of Stock Composite Index. The framework employed also proves useful for forecasting Stock Composite Index turning points. The empirical results in this paper show that ANN method is preferable to Markov-Switching model to some extent.

33.1 Introduction

Many studies conclude that stock returns can be predicted by means of macroeconomic variables with an important business cycle component. Due to the fact that the change in regime should be considered as a random event

D. Liu (✉) · L. Zhang
Department of Mathematical Sciences, Xi'an Jiaotong Liverpool University, SIP, 215123, Suzhou, China
e-mail: David.Liu@xjtlu.edu.cn

L. Zhang
e-mail: lainey3120@live.cn

L. Zhang
University of Liverpool, Liverpool, UK

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_33,
© Springer Science+Business Media B.V. 2011

and not predictable, which could motivate to analyze the Shanghai Stock Exchange Composite Index within this context. There is much empirical support that macroeconomic conditions should affect aggregate equity prices, accordingly, macroeconomic factors would be possibly used for security returns.

In order to study the dynamics of the market cycles which evolved in the Shanghai Stock Exchange Market, the Composite Index is first modeled in regime switching within a univariate Markov-Switching framework (MRS). One key feature of the MRS model is to estimate the probabilities of a specific state at a time. Past research have developed the econometric methods for estimating parameters in regime-switching models, and demonstrated how regime-switching models could characterize time series behavior of some variables, which was better than the existing single-regime models.

The concept about Markov Switching Regimes firstly dates back to "Microeconomic Theory: A Mathematical Approach" [1]. Hamilton [2] applied this model to the study of the United States business cycles and regime shifts from positive to negative growth rates in real GNP. Hamilton [2] extended Markov regime-switching models to the case of auto correlated dependent data. Hamilton and Lin also report that economic recessions are a main factor in explaining conditionally switching moments of stock market volatility [3, 4]. Similar evidences of regime switching in the volatility of stock returns have been found by Hamilton and Susmel [5], Edwards and Susmel [6], Coe [7] and [8].

Secondly, this paper deals with application of neural network method, a Radial Basis Function (RBF), on the prediction of the moving trends of the Shanghai Stock. RBFs have been employed in time series prediction with success as they can be trained to find complex relationships in the data [9].

A large number of successful applications have shown that ANN models have received considerable attention as a useful vehicle for forecasting financial variables and for time-series modeling and forecasting [10, 11]. In the early days, these studies focused on estimating the level of the return on stock price index. Current studies have reflected an interest in selecting the predictive factors as a variety of input variables to forecast stock returns by applying neural networks. Several techniques such as regression coefficients [12], autocorrelations [13], backward stepwise regression [14], and genetic algorithms [14] have been employed by researchers to perform variable subset selection [12, 13]. In addition, several researchers subjectively selected the subsets of variables based on empirical evaluations [14].

The paper is organized as follows. Section 33.2 is Data Description and Preliminary Statistics. Section 33.3 presents the research methodology. Section 33.4 presents and discusses the empirical results. The final section provides with summary and conclusion.

Table 33.1 Model summary

Model	*R*	*R* square	Adjusted *R* square	Std. error
1	0.653	0.427	0.422	768.26969
2	0.773	0.597	0.590	647.06456
3	0.834	0.695	0.688	564.83973
4	0.873	0.763	0.755	500.42457
5	0.894	0.800	0.791	461.69574

33.2 Data Description and Preliminary Statistics

33.2.1 Data Description

This paper adopts two non-linear models, Univariate Markov Switching model and Artificial Neural Network Model with respect to the behavior of Chinese Stock Exchange Composite Index using data for the period from 1999 to 2009. As Shanghai Stock Exchange is the primary stock market in China and Shanghai A Share Composite is the main index reflection of Chinese Stock Market, this research adopts the Shanghai Composite (A Share). The data consist of daily observations of the Shanghai Stock Exchange Market general price index for the period 29 October 1999 to 31 August 2009, excluding all weekends and holidays giving a total of 2369 observations. For both the MRS and the ANN models, the series are taken in natural logarithms.

33.2.2 Preliminary Statistics

In this part we will explore the relationship among Shanghai Composite and Consumer Price Index, Retail Price Index, Corporate Goods Price Index, Social Retail Goods Index, Money Supply, Consumer Confidence Index, Stock Trading by using various t-tests, and regression analysis to pick out the most relevant variables as the influence factors in our research.

By using regression analysis we test the hypothesis and identify correlations between the variables. In the following multiple regression analysis we will test the following hypothesis and see whether they hold true:

$$H_0 = \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

$$H_1 = \text{At least some of the } \beta \text{ is not equal 0 (regression insignificant).}$$

In Table 33.1, *R*-square (R^2) is the proportion of variance in the dependent variable (Shanghai Composite Index) which can be predicted from the independent variables. This value indicates that 80% of the variance in Shanghai Composite Index can be predicted from the variables Consumer Price Index, Retail Price

Table 33.2 ANOVA

Model		Sum of squares	df	Mean square	*F*	Sig.
1	Regression	5.323E7	1	5.323E7	90.178	0.000
	Residual	7.142E7	121	590238.317		
	Total	1.246E8	122			
2	Regression	7.440E7	2	3.720E7	88.851	0.000
	Residual	5.024E7	120	418692.539		
	Total	1.246E8	122			
3	Regression	8.668E7	3	2.889E7	90.561	0.000
	Residual	3.797E7	119	319043.922		
	Total	1.246E8	122			
4	Regression	9.510E7	4	2.377E7	94.934	0.000
	Residual	2.955E7	118	250424.748		
	Total	1.246E8	122			
5	Regression	9.971E7	5	1.994E7	93.549	0.000
	Residual	2.494E7	117	213162.957		
	Total	1.246E8	122			

Index, Corporate Goods Price Index, Social Retail Goods Index, Money Supply, Consumer Confidence Index, and Stock Trading. It is worth pointing out that this is an overall measure of the strength of association, and does not reflect the extent to which any particular independent variable is associated with the dependent variable.

In Table 33.2, the p-value is compared to alpha level (typically 0.05). This gives the *F*-test which is significant as p-value $= 0.000$. This means that we reject the null that Stock Trading, Consumer Price Index, Consumer Confidence Index, Corporate Goods Price Index, Money Supply have no effect on Shanghai Composite.

The p value (Sig.) from the *F*-test in ANOVA table is 0.000, which is less than 0.001, implying that we reject the null hypothesis that the regression coefficients (β's) are all simultaneously correlated.

By looking at the Sig. column in particular, we gather that Stock Trading, Consumer Price Index, Consumer Confidence Index, Corporate Goods Price Index, Money Supply are variables with p-values less than 0.02 and hence VERY significant.

Then look at Fig. 33.1, the correlation numbers measure the strength and direction of the linear relationship between the dependent and independent variables. To show these correlations visually we use partial regression plots. Correlation points tend to form along a line going from the bottom left to the upper right, which is the same as saying that the correlation is positive. We conclude that Stock Trading, Consumer Price Index, Consumer Confidence Index, Corporate Goods Price Index, Money Supply and their correlation with Shanghai Composite Index is positive because the points tend to form along this line.

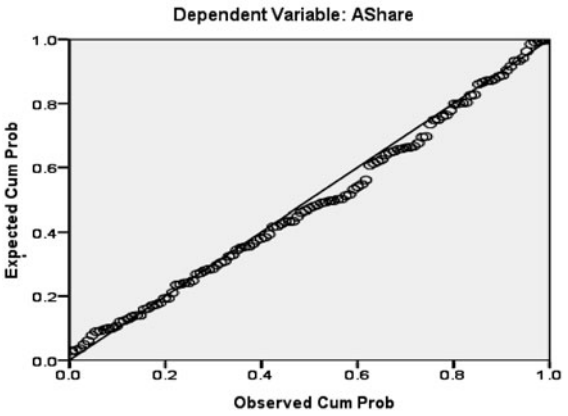


Fig. 33.1 Normal P–P plot regression standardized residual

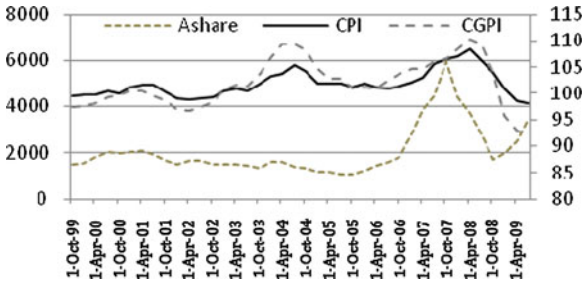


Fig. 33.2 China CPI, CGPI and Shanghai A Share Composite Index

Due to CPI Index, CGPI Index and Money Supply Increased Ratio (M1 Increased Ratio − M2 Increased Ratio) are the most correlated influence factors with Share Composite among other factors, therefore, we choose macroeconomic indicators as mentioned by Qi and Maddala [12], CPI Index, CGPI Index and Money Supply Increased Ratio (M1 Increased Ratio − M2 Increased Ratio) as well as a data set from Shanghai Stock Exchange Market are used for the experiments to test the forecasting accuracy of RBF [12]. Typically, Figs. 33.2 and 33.3 show the developments of Shanghai Composite index with CPI, CGPI and MS along time.

33.3 Empirical Models

In this section, the univariate Markov Switching Model developed by Hamilton [2] was adopted to explore regime switching of Shanghai Stock Exchange Composite Index, followed by developing an artificial neural network (ANN)—a RBF method to predict stock index moving trends. We use the RBF method to find the relationship of CPI Index, CGPI Index and Money Supply Increased Ratio with Stock Composite Index. By using the Matlab Neural Network Toolbox, RBF Network is

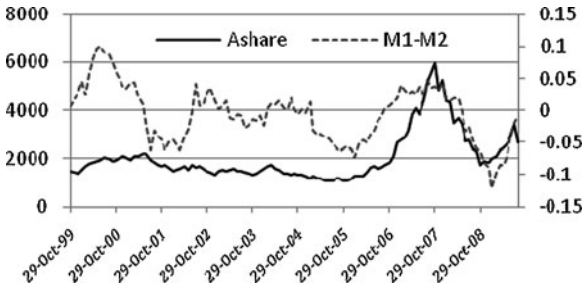


Fig. 33.3 China money supply increased (annual basis) and Shanghai A Share Composite Index

designed in a more efficient design (newrb). Finally, the forecasting performances of these two competing non-linear models are compared.

33.3.1 Markov Regime Switching Model and Estimation

33.3.1.1 Markov Regime Switching Model

The comparison of the in sample forecasts is done on the basis of the Markov Switching/Hamilton filter mathematical notation, using the Marcelo Perlin (21 June 2009 updated) forecasting modeling.

A potentially useful approach to model nonlinearities in time series is to assume different behavior (structural break) in one subsample (or regime) to another. If the dates of the regimes switches are known, modeling can be worked out with dummy variables. For example, consider the following regression model:

$$y_t = X_t'\beta_{s_t} + \varepsilon_t (t = 1, \ldots, T) \tag{33.1}$$

where, $\varepsilon_t - NID(0, \sigma_{s_t}^2), \beta_{s_t} = \beta_0(1 - S_t) + \beta_1 S_t, \ \sigma_{s_t}^2 = \sigma_0^2(1 - S_t) + \sigma_1^2 S_t, S_t = 0$ or 1, (Regime 0 or 1).

Usually it is assumed that the possible difference between the regimes is a mean and volatility shift, but no autoregressive change. That is:

$$y_t = \mu_t S_t + \phi(y_{t-1} - \mu_t S_{t-1}) + \varepsilon_t \quad \varepsilon_t - NID(0, \sigma_{s_t}^2). \tag{33.2}$$

where, $\mu_t S_t = \mu_0(1 - S_t) + \mu_1 S_t$. If $S_t (t = 1, \ldots, T)$ is known as a priori, then the problem is just a usual dummy variable auto-regression problem.

In practice, however, the prevailing regime is not usually directly observable. Denote then $P(S_t = j/S_{t-1} = i) = P_{ij}, (i, j = 0, 1)$ called transition probabilities, with $P_{i0} + P_{i1} = 1, i = 0, 1$. This kind of process, where the current state depends only on the state before, is called a Markov process, and the model a Markov switching model in the mean and the variance. The probabilities in a Markov process can be conveniently presented in matrix form:

$$\begin{pmatrix} P(S_t=0) \\ P(S_t=1) \end{pmatrix} = \begin{pmatrix} p_{00} & p_{10} \\ p_{01} & p_{11} \end{pmatrix} \begin{pmatrix} P(S_{t-1}=0) \\ P(S_{t-1}=1) \end{pmatrix}$$

Estimation of the transition probabilities P_{ij} is usually done (numerically) by maximum likelihood as follows. The conditional probability densities function for the observations y_t, given the state variables, S_{t-1} and the previous observations $F_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\}$ is

$$f(y_t/S_t, S_{t-1}, F_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_{s_t}^2}} \exp\left[-\frac{[y_t - \mu_t S_t - \phi(y_{t-1} - \mu_t S_{t-1})]^2}{2\sigma_{s_t}^2}\right] \tag{33.3}$$

$$\varepsilon_t = y_t - \mu_t S_t - \phi(y_{t-1} - \mu_t S_{t-1}) - NID(0, \sigma_{s_t}^2)$$

The chain rule for conditional probabilities yields then for the joint probability density function for the variables y_t, S_t, S_{t-1}, given past information F_{t-1}, $f(y_t, S_t, S_{t-1}/F_{t-1}) = f(y_t/S_t, S_{t-1}, F_{t-1})P(S_t, S_{t-1}/F_{t-1})$, such that the log-likelihood function to be maximized with respect to the unknown parameters becomes

$$\begin{aligned} l_t(\theta) &= \log\left[\sum_{S_t=0}^{1}\sum_{S_{t-1}=0}^{1} f(y_t/S_t, S_{t-1}, F_{t-1})P(S_t, S_{t-1}/F_{t-1})\right] \\ \theta &= (p, q, \phi, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2) \end{aligned} \tag{33.4}$$

and the transition probabilities: $p = P(S_t = 0/S_{t-1} = 0)$ and $q = P(S_t = 1/S_{t-1} = 1)$. Steady state probabilities $P(S_0 = 1/F_0)$ and $P(S_0 = 0/F_0)$ are called the steady state probabilities, and, given the transition probabilities p and q are obtained as:

$$P(S_0 = 1/F_0) = \frac{1-p}{2-q-p}, \quad P(S_0 = 0/F_0) = \frac{1-q}{2-q-p}.$$

33.3.1.2 Stock Composite Index Moving Trends Estimation

In our case, we have three explanatory variables X_{1t}, X_{2t}, X_{3t} in a Gaussian framework (Normal distribution) and the input argument S, which is equal to $S = [1111]$, then the model for the mean equation is:

$$y_t = X_{1t}\beta_{1,S_t} + X_{2t}\beta_{2,S_t} + X_{3t}\beta_{3,S_t} + \varepsilon_t \quad \varepsilon_t - NID(0, \sigma_{s_t}^2) \tag{33.5}$$

where, S_t represents the state at time t, that is, $S_t = 1, \ldots, K$ (K is the number of states); $\sigma_{s_t}^2$ is Error variance at state S_t; β_{S_t} is beta coefficient for explanatory variable i at state S_t, where i goes from 1 to n; ε_t is residual vector which follows a particular distribution (in this case Normal).

With this change in the input argument S, the coefficients and the model's variance are switching according to the transition probabilities. Therefore, the logic is clear: the first elements of input argument S control the switching dynamic

of the mean equation, while the last terms control the switching dynamic of the residual vector, including distribution parameters.

Based on Gaussian maximum likelihood, the equations are represented as following: State 1 (= 1), $y_t = X_{1t}\beta_{1,1} + X_{2t}\beta_{2,1} + X_{3t}\beta_{3,1} + \varepsilon_t$; State 2 (= 2), $y_t = X_{1t}\beta_{1,2} + X_{2t}\beta_{2,2} + X_{3t}\beta_{3,2} + \varepsilon_t$. With $\begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix}$ as the transition matrix, which controls the probability of a regime switch from state j (column j) to state i (row i). The sum of each column in P is equal to one, since they represent full probabilities of the process for each state.

33.3.2 Radial Basis Function Neural Networks

The specific type of ANN employed in this study is the Radial Basis Function (RBF), the most widely used among the many types of neural networks. RBFs were first used to solve the interpolation problem-fitting a curve exactly through a set of points. Fausett defines radial basis functions as "activation functions with a local field of response at the output" [15]. The RBF neural networks are trained to generate both time series forecasts and certainty factors.

The RBF neural network is composed of three layers of nodes. The first is the input layer that feeds the input data to each of the nodes in the second or hidden layer. The second layer of nodes differs greatly from other neural networks in that each node represents a data cluster which is centered at a particular point and has a given radius. The third and final layer consists of only one node. It acts to sum the outputs of the second layer of nodes to yield the decision value [16].

The ith neurons input of a hidden layer is $k_i^q = \sqrt{\sum_{j=1}^{n} (W1_{ji} - X_j^q)^2} \times b1_i$ and output is:

$$r_i^q = \exp(-(k_i^q)^2) = \exp\left(\sqrt{\sum_{j=1}^{n} (W1_{ji} - X_j^q)^2} \times b1_i\right)$$
$$= \exp\left(-\left(\left\|W1_i - X_j^q\right\| \times b1_i\right)^2\right)$$

where, $b1_i$ presents threshold value, X_j is the input feature vector and the approximant output r_i^q is differentiable with respect to the weights $W1_i$.

When an input vector is fed into each node of the hidden layer simultaneously, each node then calculates the distance from the input vector to its own center. That distance value is transformed via some function, and the result is output from the node. That value output from the hidden layer node is multiplied by a constant or

weighting value. That product is fed into the third layer node which sums all the products and any numeric constant inputs. Lastly, the third layer node outputs the decision value.

A Gaussian basis function for the hidden units given as Z_j for $j = 1, \ldots, J$, where

$$Z_j = \exp\left(\frac{-\left(\|X - \mu_j\|\right)^2}{2\sigma^2}\right).$$

μ_j and σ_j are mean and the standard deviation respectively, of the jth unit receptive field and the norm is Euclidean.

In order to obtain the tendency of A Share Composite Index, we examine the sample performance of quarterly returns (totally 40 quarters) forecasts for the Shanghai Stock Exchange Market from October 1999 to August 2009, using three exogenous macroeconomic variables, the CPI, CGPI and Money Supply (M1–M2, Increased on annual basis) as the inputs to the model. We use a Radial Basis Function network based on the learning algorithm presented above. Using the Matlab Neural Network Toolbox, the RBF network is created using an efficient design (newrb). According to Hagan et al. [17], a small spread constant can result in a steep radial basis curve while a large spread constant results in a smooth radial basis curve; therefore it is better to force a small number of neurons to respond to an input. Our interest goes to obtain a single consensus forecast output, the sign of the prediction only, which will be compared to the real sign of the prediction variable. After several tests and changes to the spread, at last we find spread $= 4$ is quite satisfied for out test. As a good starting value for the spread constant is between 2 and 8 [17], we set the first nine columns of y' as the test samples [17].

33.4 Empirical Results

33.4.1 Stock Composite Index Moving Trends Estimation by MRS

Table 33.3 shows the estimated coefficients of the proposed MRS along with the necessary test statistics for evaluation of Stock Composite Index Moving Trends. The Likelihood Ratio test for the null hypothesis of linearity is statistically significant and this suggests that the linearity is strongly rejected. The results in Table 33.3 further highlight several other points: First, value of the switching variable at state 1 is 0.7506, at state 2 value of the switching variable is −0.0161; and the model's standard deviation σ takes the values of 0.0893 and 0.0688 for regime 1 and regime 2 respectively; these values help us to identify regime 1 as the upward regime and regime 2 as the downward regime. Second, the duration measure shows that the upward regime lasts approximately 57 months, whereas the high volatility regime lasts approximately 24 months.

Table 33.3 Stock index moving trends estimation by MRS

Parameters	Estimate	Std err
μ_0	0.7506	0.0866
μ_2	−0.0161	0.0627
σ_0^2	0.0893	0.0078
σ_2^2	0.0688	0.0076
Expected duration	56.98 time periods	23.58 time periods
Transition probabilities		
p (regime1)		0.98
q (regime0)		0.96
Final log likelihood		119.9846

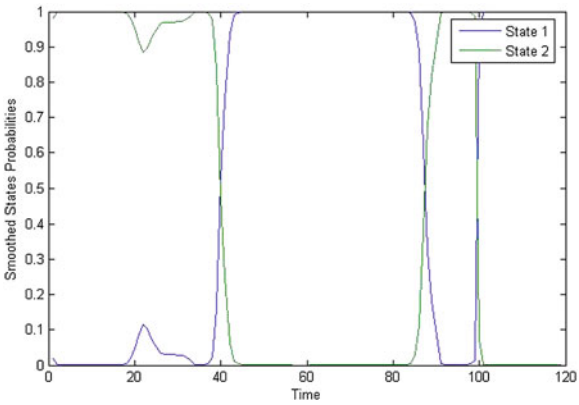


Fig. 33.4 Smoothed states probabilities (moving trends)

As we use the quarterly data for estimating the Moving Trends, the smoothed probabilities and filtered state probabilities lines seem exiguous. Figure 33.4 reveals the resulting smoothed probabilities of being in up and down moving trends regimes along Shanghai Stock Exchange Market general price index. Moreover, filtered States Probabilities is shown in Fig. 33.5, several periods of the sample are characterized by moving downwards associated with the presence of a rational bubble in the capital market of China from 1999 to 2009.

33.4.2 Radial Basis Function Neural Networks

Interestingly, the best results we obtained from RBF training are 100% correct approximations of the sign of the test set, and 90% of the series on the training set. This conclusion on one hand is consensus with the discovery in "the Stock Market and the Business Cycle" by Hamilton and Lin [3]. Hamilton and Lin [3] argued that the analysis of macroeconomic fundamentals was certainly a satisfactory explanation for stock volatility. To our best knowledge, the fluctuations in the

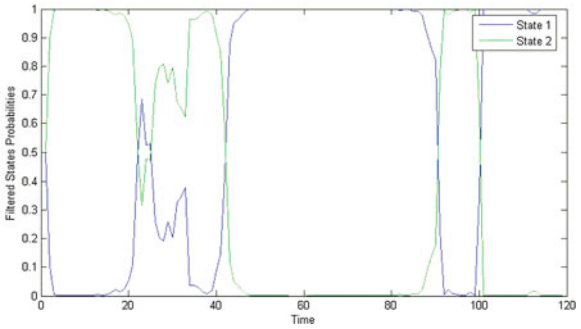


Fig. 33.5 Filtered states probabilities (moving trends)

Table 33.4 RBF training output

x	y'	T	x	y'	T
0.80937	1	1	0.031984	0	0
0.30922	0	0	0.80774	1	0
0.96807	1	1	0.68064	1	0
1.0459	1	1	0.74969	1	0
−0.011928	0	0	0.54251	1	1
0.92	1	1	0.91874	1	1
0.81828	1	1	0.50662	1	1
0.054912	0	0	0.44189	0	1
0.34783	0	0	0.59748	1	1
0.80987	1	1	0.69514	1	1
1.1605	1	1	1.0795	1	1
0.66608	1	1	0.16416	0	0
0.22703	0	0	0.97289	1	1
0.45323	0	0	−0.1197	0	0
0.69459	1	1	0.028258	0	0
0.16862	0	0	0.087562	0	0
0.83891	1	1	−0.084324	0	0
0.61556	1	1	1.0243	1	1
1.0808	1	1	0.98467	1	1
−0.089779	0	0	0.0032105	0	0

level of macroeconomic variables such as CPI and CGPI and other economic activity are a key determinant of the level of stock returns [18]. On the other hand, in a related application, also showed that RBFs have the "best" approximation property-there is always a choice for the parameters that is better than any other possible choice—a property that is not shared by MLPs.

Due to the Normal Distributions intervals, outputs is $y' = F(X)$, $F(X) = 1$ if $X \geq 0.5$, $F(X) = 0$ if $X < 0.5$.

Table 33.4 gives the results of the outputs. From x we could know that the duration of regime 1 is 24 quarters and regime 0 is 16 quarters. The comparisons of MRS and RBF models can be seen in Table 33.5. It is clear that the RBF model outperforms the MRS model on the regime duration estimation.

Table 33.5 Regime comparison of stock index moving trends

Model	Regime 1 (months)	Regime 0 (months)
Observed durations	66	54
Markov-switching	57	24
Radial basis function	72	48

33.5 Conclusion and Future Work

In this chapter, we compared the forecasting performance of two nonlinear models to address issues with respect to the behaviors of aggregate stock returns of Chinese Stock Market. Rigorous comparisons between the two nonlinear estimation methods have been made. From the Markov-Regime Switching model, it can be concluded that real output growth is subject to abrupt changes in the mean associated with economy states. On the other hand, the ANN method developed with the prediction algorithm to obtain abnormal stock returns, indicates that stock returns should take into account the level of the influence generated by macroeconomic variables. Further study will concentrate on prediction of market volatility using this research framework.

Acknowledgments This work was supported by the Pilot Funds (2009) from Suzhou Municipal Government (Singapore Industrial Park and Higher Educational Town—SIPEDI) for XJTLU Lab for Research in Financial Mathematics and Computing.

References

1. Henderson JM, Richard E (1958) Quandt, micro-economic theory: a mathematical approach. McGraw-Hill, New York
2. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57(2):357–384
3. Hamilton JD, Lin G (1996) Stock market volatility and the business cycle. J Appl Econom 11(5):573–593
4. Hamilton JD (1996) Specification tests in Markov-switching Time-series models. J Econom 70(1):127–157
5. Hamilton JD, Susmel R (1994) Autoregressive conditional heteroskedasticity and changes in regime. J Econom 64(1–2):307–333
6. Edwards S, Susmel R (2001) Volatility dependence and contagion in emerging equities markets. J Dev Econ 66(2):505–532
7. Coe PJ (2002) Financial crisis and the great depression: a regime switching approach. J Money Credit Bank 34(1):76–93
8. Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
9. Chen S, Cowan CFN, Grant PM (1991) Orthogonal least squares learning algorithm for radial basis function network. IEEE Trans Neural Netw 2(2):302–309
10. Swanson N, White H (1995) A model selection approach to assessing the information in the term structure using linear models and artificial neural networks. J Bus Econ Stat 13(8): 265–275
11. Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: the state of the art. Int J Forecast 14(1):35–62

12. Qi M, Maddala GS (1999) Economic factors and the stock market: a new perspective. J Forecast 18(3):151–166
13. Desai VS, Bharati R (1998) The efficiency of neural networks in predicting returns on stock and bond indices. Decis Sci 29(2):405–425
14. Motiwalla L, Wahab M (2000) Predictable variation and profitable trading of US equities: a trading simulation using neural networks. Comput Oper Res 27(11–12):1111–1129
15. Fausett L (1994) Fundamentals of neural networks: architectures, algorithms and applications. Prentice-Hall, Upper Saddle River
16. Moody J, Darken C (1989) Fast learning in networks of locally tuned processing units. Neural Comput 1(2):281–294
17. Hagan MT, Demuth HB, Beale MH (1996) Neural network design. PWS Publishing, Boston
18. Liu D, Zhang L (2010) China stock market regimes prediction with artificial neural network and markov regime switching. In: Lecture notes in engineering and computer science: proceeding of the world congress on engineering 2010, WCE 2010, 30 June–2 July, 2010 London, UK, pp 378–383

Chapter 34
Fund of Hedge Funds Portfolio Optimisation Using a Global Optimisation Algorithm

Bernard Minsky, M. Obradovic, Q. Tang and Rishi Thapar

Abstract Portfolio optimisation for a Fund of Hedge Funds ("FoHF") has to address the asymmetric, non-Gaussian nature of the underlying returns distributions. Furthermore, the objective functions and constraints are not necessarily convex or even smooth. Therefore traditional portfolio optimisation methods such as mean–variance optimisation are not appropriate for such problems and global search optimisation algorithms could serve better to address such problems. Also, in implementing such an approach the goal is to incorporate information as to the future expected outcomes to determine the optimised portfolio rather than optimise a portfolio on historic performance. In this paper, we consider the suitability of global search optimisation algorithms applied to FoHF portfolios, and using one of these algorithms to construct an optimal portfolio of investable hedge fund indices given forecast views of the future and our confidence in such views.

B. Minsky (✉) · R. Thapar
International Asset Management Ltd., 7 Clifford Street, London, W1S 2FT, UK
e-mail: bminsky@iam.uk.com

R. Thapar
e-mail: rthapar@iam.uk.com

M. Obradovic · Q. Tang
School of Mathematical and Physical Sciences, Sussex University, Brighton, BN1 9RF, UK
e-mail: mo32@sussex.ac.uk

Q. Tang
e-mail: q.tang@sussex.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_34,
© Springer Science+Business Media B.V. 2011

34.1 Introduction

The motivation for this paper was to develop a more robust approach to constructing portfolios of hedge fund investments that takes account of the issues that confront portfolio managers:

1. The non-Gaussian, asymmetric nature of hedge fund returns;
2. The tendency of optimisation algorithms to find corner solutions;
3. The speed in computation and efficiency in finding the solution; and
4. The desire to incorporate forecast views into the problem specification.

We describe here how each of these issues was addressed and illustrate with reference to the optimising of a portfolio of investable hedge fund indices. This paper synthesises a review of the applicability of global search optimisation algorithms for financial portfolio optimisation with the development of a Monte Carlo simulation approach to forecasting hedge fund returns and implementing the methodology into an integrated forecasting and optimisation application.

In Sect. 34.2, we summarise the review of global search optimisation algorithms and their applicability to the FoHF portfolio optimisation problem. In Sect. 34.3, we describe the Monte Carlo simulation technique adopted using resampled historical returns data of hedge fund managers and also how we incorporated forecast views and confidence levels, expressed as probability outcomes, into our returns distribution data. In Sect. 34.4, we report the results of applying the methodology to a FoHF portfolio optimisation problem and in Sect. 34.5, we draw our conclusions from the study.

34.2 Review of Global Search Optimisation Algorithms

The FoHF portfolio optimisation problem is an example of the typical minimisation problem in finance:

$$\begin{array}{l} \min \\ g(x)(<) = g0 \\ \vdots \\ h(x)(<) = h0 \end{array} f(x)$$

where f is non-convex and maybe non-smooth, called the objective function. The g, …, h are constraint functions, with $g0$, …, $h0$ as minimum thresholds. The variable x usually denotes the weights assigned to each asset and the constraints will usually include the buying and shorting limits on each asset.

It is well known that many of the objective functions and constraints specified in financial minimisation problems are not differentiable. Traditional asset management has relied on the Markowitz specification as a mean–variance optimisation problem which is soluble by classical optimisation methods. However,

in FoHF portfolio optimisation the distribution of hedge fund returns are non-Gaussian and the typical objective functions and constraints are not limited to simple mean, variance and higher order moments of the distribution. We have previously [1] discussed the use of performance and risk statistics such as maximum drawdown, downside deviation, co-drawdown, and omega as potential objective functions and constraint functions which are not obviously differentiable.

With the ready availability of powerful computing abilities and less demand on smoothness, it is possible to look for global optimisation algorithms which do not require regularity of the objective (constraint) functions to solve the financial minimisation problem.

In our review of the literature [1], we found that there are three main ideas of global optimisation; Direct, Genetic Algorithm, and Simulated Annealing. In addition, there are a number of other methods which are derived from one or more of the ideas listed above.

A key characteristic of fund of fund portfolio optimisation, in common with other portfolio optimisation problems, is that the dimensionality of the problem space is large. Typically, a portfolio of hedge funds will have between 20 and 40 assets with some commingled funds having significantly more assets. This means that the search algorithm cannot conduct an exhaustive test of the whole space efficiently. For example, if we have a portfolio of 40 assets we have a 40-dimensional space, and an initial grid of 100 points on each axis produces 1040 initial test points to evaluate the region where the global minimum might be found. This would require considerable computing power and would not be readily feasible.

Each of the methods we considered in our review requires an initial search set. The choice of the initial search set is important as the quality of the set impacts the workload required to find the global minimum. The actual approach to moving from the initial set to finding better and better solutions differs across the methods and our search also revealed some approaches that combine the methods to produce a hybrid algorithm.

In the paper [1] we evaluated seven algorithms across the methods to identify which method and specific algorithm was best suited to our FoHF portfolio optimisation problem. The algorithms considered are described here.

34.2.1 PGSL: Probabilistic Global Search Lausanne

PGSL is a hybrid algorithm, proposed by Raphael [2], drawing on the Simulated Annealing method that adapts its search grid to concentrate on regions in the search space that are favourable and to intensify the density of sampling in these attractive regions.

The search space is sampled using a probability distribution function for each axis of the multi-dimensional search space. At the outset of the search process, the probability distribution function is a uniform distribution with intervals of

constant width. During the process, a probability distribution function is updated by increasing probability and decreasing the width of intervals of the regions with good functional values. A focusing algorithm is used to progressively narrow the search space by changing the minimum and maximum of each dimension of the search space.

34.2.2 MCS: Multi-Level Co-ordinate Search

MCS belongs to the family of branch and bound methods and it seeks to solve bound constrained optimisation problems by combining global search (by partitioning the search space into smaller boxes) and local search (by partitioning sub-boxes based on desired functional values). In this way, the search is focused in favour of sub-boxes where low functional values are expected. The balance between global and local parts of the search is obtained using a multi-level approach. The sub-boxes are assigned a level, which is a measure of how many times a sub-box has processed. The global search part of the optimisation process starts with the sub-boxes that have low level values. At each level, the box with lowest functional value determines the local search process. The optimisation method is described in the paper by Huyer and Neumaier [3].

Some of the finance papers that have examined MCS include Aggregating Risk Capital [4] and Optimising Omega [5]. In Optimising Omega [5] Value-at-Risk of a portfolio is calculated using marginal distributions of the risk factors and MCS is employed to search for the best-possible lower bound on the joint distribution of marginal distributions of the risk factors. Optimising Omega [5] uses MCS to optimise for Omega ratio, a non-smooth performance measure, of a portfolio.

34.2.3 MATLAB Direct

The Direct Search algorithm, available in MATLAB's Genetic Algorithm and Direct Search Toolbox, uses a pattern search methodology for solving bound linear or non-linear optimisation problems [6]. The algorithms used are Generalised Pattern Search (GPS) and Mesh Adaptive Search (MADS) algorithm.

The pattern search algorithm generates a set of search directions or search points to approach an optimal point. Around each search point, an area, called a mesh, is formed by adding the current point to a scalar multiple of a set of vectors called a pattern. If a point in the mesh is found that improves the objective function at the current point, the new point becomes the current point for the next step and so on. The GPS method uses fixed direction vectors and MADS uses random vectors to define a mesh.

34.2.4 MATLAB Simulated Annealing

The Simulated Annealing method uses probabilistic search algorithm models that model the physical process of heating a material and then slowly lowering the temperature to decrease defects, thus minimising the system energy [6]. By analogy with this physical process, each step in the Simulated Annealing algorithm replaces the current point by another point that is chosen depending on the difference between the functional values at the two points and the temperature variable, which is systematically decreased during the process.

34.2.5 MATLAB Genetic Algorithm

The MATLAB's Genetic Algorithm is based on the principles of natural selection and uses the idea of mutation to produce new points in the search for an optimised solution [6]. At each step, the Genetic Algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. In this way, the population evolves toward an optimal solution.

34.2.6 TOMLAB LGO

Tomlab's Global Optimiser, TOMLAB/LGO, combines global and local search methodologies [7]. The global search is implemented using the branch and bound method and adaptive random search. The local search is implemented using a generalised reduced gradient algorithm.

34.2.7 NAG Global Optimiser

NAG's Global Optimiser, E05JBF, is based on MCS, as described above. E05JBF is described in NAG's Library Routine Document [8] and Optimising Omega [5].

The above algorithms were evaluated on the three constrained optimisation problems. The constraints consisted of both linear constraints on the allocation weights to the assets and constraints on the level of functions that characterise the portfolio's performance or risk. The algorithms were measured regarding time to run, percentage of corners in the optimal solution, and the deviation from the average optimal solution. A simple scoring rule combining these three factors as a weighted sum was constructed.

There was considerable variation in relative performance between the algorithms across the different tests. Two algorithms, MATLAB Annealing and

MATLAB Genetics, were found to be unstable giving rise to different results when repeated runs of the same problem and environment were performed. They also produced widely different results, from very good to very bad, across the tests and were rejected from consideration easily. The other five algorithms all produced acceptable results with MATLAB Direct scoring best across the constrained optimisation examples. PGSL, the adaptive Simulated Annealing algorithm performs reasonably in most tests and has been used by IAM for the past 4 years. Therefore, we chose to compare MATLAB Direct with PGSL in our portfolio optimisation implementation.

34.3 Implementing the Global Search Optimisation Algorithm

Traditional optimisation of portfolios has focused on determining the optimal portfolio given the history of asset returns and assuming that the distribution of returns is Gaussian and stationary over time. Our experience is that these assumptions do not hold and that any optimisation should use the best forecast we can make of the horizon for which the portfolio is being optimised. When investing in hedge funds, liquidity terms are quite onerous with lock ups and redemption terms from monthly to annual frequencies, and notice periods ranging from a few days to 6 months. This means that the investment horizon tends to be 6–12 months ahead to reflect the minimum time any investment will be in a portfolio.

The forecast performance of the assets within the portfolio is produced using Monte Carlo simulation and re-sampling. The objective is to produce a random sample of likely outcomes period by period for the forecast horizon based on the empirical distributions observed for the assets modified by our views as to the likely performance of the individual assets. This is clearly a non-trivial exercise, further complicated by our wish to maintain the relationship between the asset distributions and any embedded serial correlation within the individual asset distribution.

The approach implemented has three components:

1. Constructing a joint distribution of the asset returns from which to sample;
2. Simulating the returns of the assets over the forecast horizon; and
3. Calculating the relevant objective function and constraints for the optimisation.

34.3.1 Constructing the Joint Distribution of Asset Returns

We used bootstrapping in a Monte-Carlo simulation framework to produce the distribution of future portfolio returns. Bootstrapping is a means of using the available data by resampling with replacement. This generates a richer sample than would otherwise be available. To preserve the relationship between the assets we

treat the set of returns for the assets in a time period as an observation of the joint distribution of the asset returns. An enhancement to this sampling scheme to capture any serial correlation is to block sample a group contiguously, say three periods together. Block sampling of three periods at a time offers around 10 million distinct samples of blocks of three time periods.

As we used bootstrapping to sample from the distribution and we wished to preserve the characteristics of the joint distribution, we needed to define a time range over which we have returns for all of the assets in the portfolio. Hedge funds report returns generally on a monthly basis, which means that we needed to go back a reasonable period of time to obtain a sufficiently large number of observations to enable the bootstrap sampling to be effective. For hedge funds this is complicated because many of the funds have not been in existence for very long, with the median life of a hedge fund being approximately 3 years. Although the longer the range that can be used for the joint distribution the greater the number of points available for sampling, the lack of stationarity within the distribution leads us to select a compromise period, typically 5 years, as the desired range. Where a hedge fund does not have a complete 5 year history, we employed a backfill methodology to provide the missing data.

There are a number of approaches to backfilling asset return time series such as selecting a proxy asset to fill the series; using a strategy index with a random noise component; constructing a factor model of the asset returns from the available history and using the factor return history and model to backfill; or to randomly select an asset from a set of candidate assets that could have been chosen for the portfolio for the periods that the actual asset did not exist. We adopted this last method, selecting an asset from a set of available candidates within a peer group for the missing asset. Where the range for which returns are missing was long, we repeated the exercise of selecting an asset at random from the available candidates within the peer group, say, every six periods. Our reasoning for applying this approach is that we assume as portfolio managers, given the strategy allocation of the portfolio, that we would have chosen an asset from the candidate peer group available at that time to complete the portfolio. Using this process we constructed a complete set of returns for each of the assets going back, say, 5 years. The quality of the backfill depends on how narrowly defined the candidate peer group is defined. At International Asset Management Limited (IAM), we have defined our internal set of strategy peer groups that reflect best our own interpretation of the strategies in which we invest. This is because hedge fund classifications adopted by most of the index providers tend to be broad, and can include funds that would not feature in IAM's classifications.

34.3.2 Simulating Returns Over the Forecast Horizon

We simulated the returns of the assets using a block bootstrap of the empirical joint distributions, which are modified by probabilistically shifting the expected

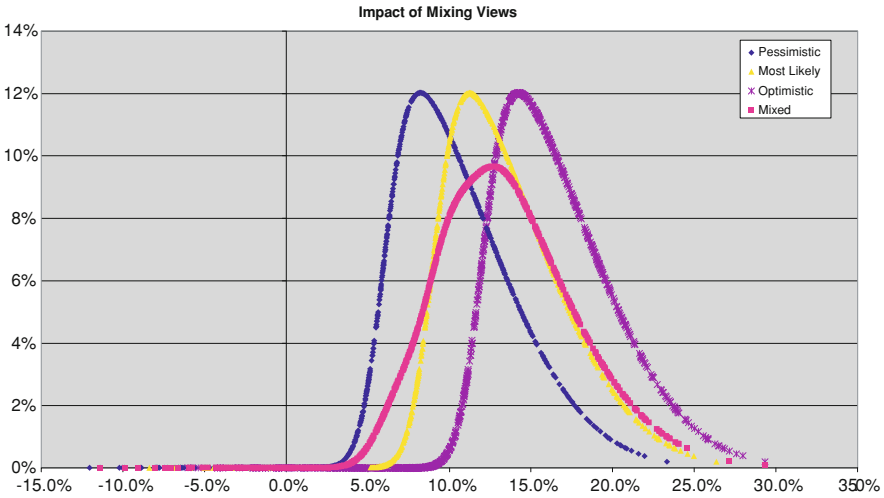


Fig. 34.1 Probabilistic shifting of expected mean

return of the sample according to our assessment of the likely return outcomes for the assets. First we describe the process of incorporating forecast views by expectation shifting and then we describe the block bootstrapping method.

The desire to include forecast views, expressed as expected annual returns, and confidence, expressed as probabilities, within a portfolio optimisation problem has been addressed in a number of ways. Black and Litterman developed an approach where the modeller expressed a view as to the expected mean of a returns series and attached a confidence to each view. This approach is Bayesian and allows the traditional Mean–Variance approach to be adapted to allow for more stable and intuitive allocations which do not favour corner solutions. However, we have chosen an empirical approach, of mixing probabilistically mean shifted versions of the empirical distribution, to include views that allows a range of outcomes to be specified with a confidence associated with the views.

Figure 34.1 shows how applying a probabilistic shift to the mean of a distribution not only repositions the distribution but changes the higher order moments as the spread, skew and kurtosis all change.

In Table 34.1 the forecast views for a number of strategies are set out with associated confidence. The optimistic, pessimistic and most likely views are the best assessment of the potential expected return of the mean fund within the strategy. The confidence level represents the likelihood of that view prevailing. We note the sum of the three confidence levels is one. We use these likelihoods to determine for each asset, according to its strategy, which shift should be applied to the distribution for that simulation. This is implemented by simply sampling from the uniform distribution and dividing the distribution into three segments according to the confidence levels associated with the three views. Recognising that each asset does not track its strategy with certainty we calculate the beta for

Table 34.1 Forecast views and confidence by hedge fund strategy

Strategy	Opt. view (%)	Conf./prob. (%)	Pess. view (%)	Conf./prob. (%)	Most likely view (%)	Conf./prob. (%)
Convertible bond arbitrage	17.5	25	7.5	25	12.5	50
Credit	17.5	25	7.5	25	12.5	50
Event driven	10.0	25	0.0	25	5.0	50
Fixed income rel val	15.0	25	10.0	25	12.5	50

the asset with respect to the strategy and adjust the return by the randomly chosen shift ("k") multiplied by the asset beta calculated. So the return in any period ("t") for an asset ("a") which follows strategy ("s") for simulation trial ("n") is:

$$m\_r_{a,t,n} = \text{raw}\_r_{a,t,n} + \beta x \, \text{shift}_{s,k}$$

34.3.3 Calculating the Objective Function and Constraint Functions

In implementing the bootstrapped Monte Carlo simulation we simulate 500 trials or scenarios for the assets in the portfolio. This produces a distribution of returns of each asset and the distributions of any statistics we may wish to compute. Our objective and constraint functions are statistics based on the distribution of portfolio returns. With a set of asset allocation weights, the distribution of portfolio returns and statistics distributions may be calculated. It is worth discussing how we use this information within the optimisation algorithm. To do this we shall use as an example maximising expected return subject to a maximum level of maximum drawdown.

As we have chosen to optimise expected return, our objective function is simply the median of the distribution of portfolio returns. If we set our objective to ensure performance is at an acceptable level in most circumstances we might choose the bottom five percentile of return as the objective function so as to maximise the least likely (defined as fifth percentile) return. This reflects the flexibility we have with using a simulated distribution as the data input into the optimisation process.

In PGSL, as with almost all of the global search optimisation algorithms, both the linear and non-linear constraints are defined as penalty functions added to the objective function and hence are soft constraints rather than hard constraints that must be satisfied. The weight attached to each penalty function determines how acceptable a constraint violation is. In our example, we define the penalty function as the average of the maximum drawdown for the lowest five percentile of the maximum drawdown distribution less the constraint boundary assuming the conditional average exceeds the constraint level multiplied by an importance factor:

Table 34.2 FoHF portfolio optimisation problem

Objective	
Maximise median portfolio return	
Subject to:	
Maximum drawdown over forecast period	*Less than 5%*
Total allocations for full investment	100%
Cash	10%
Within the following constraints	
RBC Hedge 250 Equity Market Neutral	Between 10 and 16%
RBC Hedge 250 Equity Long/Short Directional	Between 14 and 20%
All Long/Short Equity	Between 24 and 36%
RBC Hedge 250 Fixed Income Arbitrage	Between 7 and 13%
RBC Hedge 250 Macro	Between 10 and 20%
RBC Hedge 250 Managed Futures	Between 10 and 20%
RBC Hedge 250 Credit	Between 5 and 15%
RBC Hedge 250 Mergers and Special Situations	Between 0 and 10%
RBC Hedge 250 Multi-Strategy	Between 0 and 10%

$$\text{Max\_dd\_penalty} = -\text{Max}(\text{Constraint\_dd} - \text{average}(\text{Max\_ddn}|\text{Lower } 5\%\text{ile}), 0)/ \\ \text{No. of Trials} * \text{Importance}$$

This measure is analogous to an expected tail loss or Conditional VaR (CVar) in that it is an estimate of the conditional expectation of the maximum drawdown for the lower tail of the distribution of drawdowns.

34.4 Results of Optimising a FoHF Portfolio

The approach to optimising a FoHF portfolio has been implemented in MATLAB and applied to a portfolio of eight RBC Hedge 250 hedge fund strategy indices. The monthly returns for indices from July 2005 are available from the RBC website. As the simulation requires 5 years of monthly returns the series were backfilled from the IAM's pre-determined group of candidate assets within the relevant investment strategy peer group, using random selection as previously described.

The portfolio was optimised with an objective function to maximise median returns subject to constraints on the maximum and minimum allocations to each asset, a constraint on the maximum and minimum allocation to Long/Short Equity strategies and a maximum allowable maximum drawdown of 5% over the forecast horizon. Thus the optimisation problem is as set out in Table 34.2.

First we noted that the total allocations satisfying the equality constraint of all capital is deployed with both PGSL and MATLAB Direct and that all the asset allocation constraints are satisfied including the constraint on all Long/Short Equity strategies by MATLAB Direct, but not by PGSL. Secondly we noted that

Table 34.3 Optimal allocations and results

Asset	Lower bound (%)	Upper bound (%)	Naïve (%)	PGSL (%)	Direct (%)
Cash	10	10	10.0	10.0	10.0
RBC Hedge 250 Equity Market Neutral	10	16	13.0	15.0	16.0
RBC Hedge 250 Equity Long/ Short	14	20	17.0	16.1	20.0
RBC Hedge 250 Fixed Income Arbitrage	7	13	10.0	12.5	13.0
RBC Hedge 250 Macro	10	20	15.0	12.7	13.9
RBC Hedge 250 Managed Futures	10	20	15.0	12.7	15.4
RBC Hedge 250 Credit	5	15	10.0	16.7[a]	11.5
RBC Hedge 250 Mergers and Sp.Situations	0	10	5.0	0.6	0.0
RBC Hedge 250 Multi-Strategy	0	10	5.0	5.8	0.1
Median return	–	–	7.40	7.80	7.96
Excess Tail Maximum Drawdown	–	–	2.70	3.22	1.71

[a] In breach of upper allocation constraint

with PGSL only one other allocation is near its lower or upper bounds whereas with MATLAB Direct five allocations are at or near either the lower or upper bounds. Thirdly we compared the results to a portfolio where the allocation of capital to the different assets was chosen to be the midpoint between the lower and upper bounds placed on each asset (the naïve allocation). We noted that both optimisers improved median returns (7.8 and 8.0% vs. 7.40%) and that MATLAB Direct reduced the breach of the maximum drawdown constraint (1.71% vs. 2.70%) whereas the PGSL optimisation failed to improve on this condition (3.22% vs. 2.70%). MATLAB Direct portfolio had the better maximum drawdown distribution both in terms of worst case and general performance. Also, MATLAB Direct optimised portfolio performs the best of the three portfolios in terms of cumulative returns. Finally, we noted that PGSL optimisation terminated on maximum iterations and this might explain why it failed to meet all the allocation criteria (Table 34.3).

34.5 Conclusion

The review of Global Search Optimisation algorithms showed that there is a range of methods available, but their relative performance is variable. The specifics of the problem and initial conditions can impact the results significantly. In applying MATLAB Direct and PGSL to the FoHF portfolio optimisation problem, we observed that we improved on the naïve solution in both cases, but each method

presented solution characteristics that might be less desirable. PGSL was unable to find a solution that met its threshold stopping criterion whilst MATLAB Direct found a solution with many corner points. Further research studies are required to evaluate the stability of the optimiser outputs and sensitivity analysis of salient optimisation parameters.

References

1. Minsky B, Obradovic M, Tang Q, Thapar R (2008) Global optimisation algorithms for financial portfolio optimisation. Working paper, University of Sussex
2. Raphael B, Smith IFC (2003) A direct stochastic algorithm for global search. Appl Math Comput 146:729–758
3. Huyer W, Neumaier A (1999) Global optimisation by multilevel coordinate search. J Global Optim 14:331–355
4. Embrechts P, Puccetti G (2006) Aggregating risk capital, with an application to operational risk. Geneva Risk Insur 31(2):71–90
5. Kane SJ, Bartholomew-Biggs MC (2009) Optimising omega. J Global Optim 45(1)
6. Genetic Algorithm and Direct Search Toolbox™ 2 user's guide, Mathworks. http://www.mathworks.com/access/helpdesk/help/pdf_doc/gads/gads_tb.pdf
7. User's guide for TOMLAB/LGO1TOMLAB. http://tomopt.com/docs/TOMLAB_LGO.pdf
8. NAG Library Routine Document E05JBF, NAG. http://www.nag.co.uk/numeric/FL/nagdoc_fl22/pdf/E05/e05jbf.pdf

Chapter 35
Increasing the Sensitivity of Variability EWMA Control Charts

Saddam Akber Abbasi and Arden Miller

Abstract Control chart is the most important statistical process control (SPC) tool used to monitor reliability and performance of manufacturing processes. Variability EWMA charts are widely used for the detection of small shifts in process dispersion. For ease in computation all the variability EWMA charts proposed so far are based on asymptotic nature of control limits. It has been shown in this study that quick detection of initial out-of-control conditions can be achieved by using exact or time varying control limits. Moreover the effect of fast initial response (FIR) feature, to further increase the sensitivity of variability EWMA charts for detecting process shifts, has not been studied so far in SPC literature. It has been observed that FIR based variability EWMA chart is more sensitive to detect process shifts than the variability charts based on time varying or asymptotic control limits.

35.1 Introduction

Control charts, introduced by Walter A. Shewhart in 1920s are the most important statistical process control (SPC) tool used to monitor the reliability and performance of manufacturing processes. The basic purpose of implementing control chart procedures is to detect abnormal variations in the process parameters

S. A. Abbasi (✉) · A. Miller
Department of Statistics, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
e-mail: sabb025@aucklanduni.ac.nz

A. Miller
e-mail: miller@stat.auckland.ac.nz

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_35,
© Springer Science+Business Media B.V. 2011

(location and scale). Although first proposed for the manufacturing industry, control charts have now been applied in a wide variety of disciplines, such as nuclear engineering [1], health care [2], education [3] and analytical laboratories [4, 5]. Shewhart-type control charts are the most widely used: process location is usually monitored by an $\overline{X}$ chart and process dispersion by an R or S chart. Research has shown that, due to the memoryless nature of Shewhart control charts, they do not perform well for the detection of small and moderate shifts in the process parameters. When quick detection of small shifts is desirable, cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) charts are superior alternatives to Shewhart charts (for details see [6, 7]).

Since the introduction of EWMA chart by [8], many researchers have examined these charts from different perspectives—see for example [5, 9–16] and the references therein. In contrast to Shewhart type charts which are only based on information of the current observations, EWMA charts make use of information from historical observations as well by adopting a varying weight scheme: the highest weight is assigned to the most recent observations and the weights decreasing exponentially for less recent observations. This helps in the earlier detection of small shifts in process (location and scale) parameters (see [6]). Monitoring process variability using EWMA chart has also attracted the attention of many researchers. Some important contributions are [17–22]. Recently [15] proposed a new EWMA chart for monitoring process dispersion, the NEWMA chart, and showed that the NEWMA chart outperformed the variability EWMA chart proposed by [19] in terms of average run length.

All the variability EWMA schemes proposed so far are based on asymptotic nature of control limits. Ease of computation has been reported as the main reason for using asymptotic limits but this makes the EWMA chart insensitive to start up quality problems. It should be noted that the exact control limits of the EWMA charts vary with time and approach the asymptotic limits as time increases (see [6]). When the process is initially out-of-control, it is extremely important to detect the sources of these out-of-control conditions as early as possible so that corrective actions can be taken at an early stage. This can be achieved by using the exact limits instead of the asymptotic control limits. The sensitivity of time varying EWMA chart can be increased further by narrowing the time varying limits at process startup or adding a head start feature. In SPC framework this feature is well known as fast initial response (FIR) (for details see [6]). The effect of FIR feature for increasing the sensitivity of variability EWMA charts has not been investigated so far in SPC literature. This study investigates the performance of variability EWMA charts that use asymptotic, time varying and FIR based control limits. The comparison has been made on the basis of run length characteristics such as average run length (ARL), median run length (MDRL) and standard deviation of run length distribution (SDRL).

To investigate the effect of time varying control limits and of FIR on variability EWMA chart performance, we use the NEWMA chart which was recently proposed by [15] in Journal of Quality Technology. Time varying and FIR based

control limits are constructed for the NEWMA chart and their performance is compared to that of asymptotic control limits. The rest of the study is organized as follows: Sect. 35.2 briefly introduces structure of the NEWMA chart and further presents the design of the NEWMA chart using time varying control limits (TNEWMA chart). The next section compares run length characteristics of NEWMA and TNEWMA charts. The effect of FIR feature is then investigated and compared to asymptotic and time varying EWMA schemes. To get a better insight on the run length distribution of these charts, run length curves are also presented. The chapter finally ends with concluding remarks.

35.2 TNEWMA Chart

In this section we briefly describe the structure of NEWMA chart as was proposed by [15] and construct time varying control limits for this chart.

Assume the quality variable of interest X follows a normal distribution with mean μ_t and variance σ_t^2 (i.e. $X \sim N(\mu_t, \sigma_t^2)$). Let S_t^2 represents the sample variance and δ_t represents the ratio of process standard deviation σ_t and its true value σ_0 at time period t (i.e. $\delta_t = \sigma_t/\sigma_0$). Suppose $Y_t = \ln(S_t^2/\sigma_0^2)$, for an in-control process i.e. $\sigma_t = \sigma_0$, Y_t is approximately normally distributed with mean μ_Y and variance σ_Y^2 where

$$\mu_Y = \ln(\delta_t^2) - \frac{1}{n-1} - \frac{1}{3(n-1)^2} + \frac{2}{15(n-1)^4} \tag{35.1}$$

and

$$\sigma_Y^2 = \frac{2}{n-1} + \frac{2}{(n-1)^2} + \frac{4}{3(n-1)^3} - \frac{16}{15(n-1)^5}. \tag{35.2}$$

Note that when the process is in control the statistic

$$Z_t = \frac{Y_t - \mu_Y | \sigma_t = \sigma_0}{\sigma_Y} \tag{35.3}$$

is exactly a standard normal variate. When the process is out of control, $Z_t \sim N(\gamma_t, 1)$, where $\gamma_t = \left(\ln(\sigma_t^2/\sigma_0^2)\right)/\sigma_Y$ [15]. The EWMA statistic for monitoring process variability used by [15] is based on resetting Z_t to zero whenever its value becomes negative i.e. $Z_t^+ = \max(0, Z_t)$. The NEWMA chart is based on plotting the EWMA statistic

$$W_t = \lambda\left(Z_t^+ - \frac{1}{2\pi}\right) + (1-\lambda)W_{t-1}, \tag{35.4}$$

where the smoothing constant λ is the weight assigned to most recent sample observation ($0 \le \lambda \le 1$). Small values of λ are effective for quick detection of small

process shifts. As the value of λ increases the NEWMA chart performs better for the detection of large process shifts. An out of control signal occurs whenever $W_t > UCL_a$ where

$$UCL_a = L_a\sqrt{\frac{\lambda}{2-\lambda}}\sigma_{Z_t^+}. \tag{35.5}$$

Ref. [23] showed that

$$\sigma^2_{Z_t^+} = \left(\frac{1}{2} - \frac{1}{2\pi}\right). \tag{35.6}$$

We will see that the exact variance of W_t is time varying and hence the exact control limit should be dependent on time approaching UCL_a as $t \to \infty$. By Defining $Z'_t = Z_t^+ - \frac{1}{2\pi}$, we can write W_t as

$$W_t = \lambda Z'_t + (1-\lambda)W_{t-1}. \tag{35.7}$$

By continuous substitution of $W_{t-i},\ i = 1, 2, \ldots, t$; the EWMA statistic W_t can be written as (see [6, 8]):

$$W_t = \lambda\sum_{i=0}^{t-1}(1-\lambda)^i Z'_{t-i} + (1-\lambda)^t W_0. \tag{35.8}$$

Taking the variance of both sides, we obtain

$$Var(W_t) = \lambda^2\sum_{i=0}^{t-1}(1-\lambda)^{2i}Var(Z'_{t-i}) + (1-\lambda)^{2t}Var(W_0). \tag{35.9}$$

For independent random observations Z'_t, $var(Z'_t) = var(Z'_{t-i}) = \sigma^2_{Z_t^+}$. After a bit of simplification, we have

$$Var(W_t) = \sigma^2_{Z_t^+}\left(\lambda^2\left[\frac{1-(1-\lambda)^{2t}}{1-(1-\lambda)^2}\right]\right). \tag{35.10}$$

This further simplifies to

$$Var(W_t) = \sigma^2_{Z_t^+}\left(\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2t}\right]\right). \tag{35.11}$$

For the rest of study we will refer to the variability EWMA chart based on exact variance of W_t given in Eq. 35.11 as the TNEWMA chart. The TNEWMA chart gives an out of control signal whenever $W_t > UCL_t$, where

$$UCL_t = L_t\sqrt{\frac{\lambda[1-(1-\lambda)^{2t}]}{2-\lambda}}\sigma_{Z_t^+}. \tag{35.12}$$

UCL_t converges to UCL_a as $t \to \infty$, where the rate of convergence is slower for smaller values of λ.

35.3 Comparison of Run Length Characteristics of NEWMA and TNEWMA Charts

To evaluate the performance of control charts, the average run length (ARL) is the most important and widely used measure. ARL indicates the mean number of observations until an out of control signal is detected by a control chart. In this study, a Monte Carlo simulation with 10,000 iterations is used to approximate run length distributions of the NEWMA and TNEWMA charts following the methods of [9, 24, 25, 26]. Note that [27, 28] indicates that even 5,000 replications are enough for finding ARLs in many control chart settings with in an acceptable error rate. To get a better insight of the performance of the proposed charts, the median and the standard deviation of the run length distribution are also provided. The summary of the run length characteristics of NEWMA and TNEWMA charts is reported in Tables 35.1 and 35.2 for different values of smoothing parameter λ. In the following tables ARL denotes the average run length, SDRL denotes the standard deviation of the run length distribution and MDRL denotes the median of the run length distribution. In each table, smoothing constant λ increases as we move across columns from left to right where as shift δ increases as we move across rows from top to bottom. The rows corresponding to $\delta = 1$ provides the run length characteristics of both charts when the process is assumed to be in statistical control. The process is said to be out-of-control for $\delta > 1.0$. Control chart multiples L_a and L_t are so chosen as to give the same in control average run length of 200 (i.e. $ARL_0 = 200$) for both the charts.

The results in Tables 35.1 and 35.2 indicate that for smaller values of λ (which is most popular choice for EWMA charts), the out-of-control ARL (ARL_1) of the TNEWMA chart is significantly lower than the ARL_1 of NEWMA chart, see for example $ARL_1 = 9.93$ for TNEWMA chart using $\lambda = 0.05$ and $\delta = 1.2$ while for NEWMA chart $ARL_1 = 14.52$ for same values of λ and δ. It indicates that TNEWMA chart requires on average nearly five less observations as compared to NEWMA chart to detect a shift of 1.2σ in process variability when $\lambda = 0.05$. MDRL of the TNEWMA chart is also lower than MDRL of the NEWMA chart while there is a slight increase in SDRL of the TNEWMA chart as compared to NEWMA chart for lower values of λ and δ. Figure 35.1 presents ARL comparison of NEWMA and TNEWMA charts for some choices of λ. In each plot, the size of multiplicative shift in process variability δ is plotted on horizontal axis while ARL is plotted on vertical axis in logarithmic scale for better visual comparison. The effect of using time varying control limits can be clearly seen from Fig. 35.1, particularly for smaller values of λ. As expected, ARL of TNEWMA chart starts to converge

Table 35.1 Run length characteristics of NEWMA chart when $ARL_0 = 200$

δ	L_a	λ									
		0.05	0.10	0.15	0.20	0.25	0.30	0.50	0.70	0.90	1.00
		1.569	*1.943*	*2.148*	*2.271*	*2.362*	*2.432*	*2.584*	*2.650*	*2.684*	*2.693*
1.0	**ARL**	**199.69**	**200.74**	**200.24**	**199.80**	**199.23**	**200.39**	**199.88**	**199.82**	**199.11**	**199.52**
	MDRL	136.00	140.50	139.00	137.00	138.00	142.00	139.00	139.00	141.00	137.00
	SDRL	197.62	198.83	200.98	197.14	197.55	203.09	197.02	196.87	195.46	202.39
1.1	**ARL**	**31.68**	**35.33**	**37.69**	**40.11**	**41.38**	**43.26**	**49.42**	**54.11**	**61.16**	**65.19**
	MDRL	24.00	26.00	28.00	29.00	29.00	31.00	35.00	38.00	43.00	46.00
	SDRL	26.26	30.77	34.30	36.79	38.99	41.50	48.52	55.14	59.77	64.62
1.2	**ARL**	**14.52**	**14.81**	**15.48**	**15.97**	**16.48**	**17.30**	**19.66**	**22.22**	**25.87**	**28.44**
	MDRL	12.00	12.00	12.00	12.00	12.00	13.00	14.00	16.00	18.00	20.00
	SDRL	10.05	11.06	12.12	13.13	13.93	15.41	18.54	21.10	25.47	27.88
1.3	**ARL**	**9.21**	**9.06**	**9.07**	**9.25**	**9.34**	**9.57**	**10.34**	**11.73**	**13.89**	**14.97**
	MDRL	8.00	8.00	7.00	7.00	7.00	7.00	8.00	8.00	10.00	10.00
	SDRL	5.49	5.87	6.43	6.83	7.35	7.55	8.89	11.04	13.51	14.43
1.4	**ARL**	**6.72**	**6.53**	**6.45**	**6.42**	**6.46**	**6.47**	**6.76**	**7.31**	**8.46**	**9.23**
	MDRL	6.00	6.00	5.00	5.00	5.00	5.00	5.00	5.00	6.00	7.00
	SDRL	3.67	3.92	4.09	4.26	4.54	4.71	5.63	6.43	7.78	8.85
1.5	**ARL**	**5.30**	**5.18**	**4.98**	**4.87**	**4.81**	**4.78**	**4.86**	**5.15**	**5.86**	**6.38**
	MDRL	5.00	5.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	5.00
	SDRL	2.69	2.83	2.89	2.93	3.12	3.26	3.74	4.32	5.22	5.93
1.6	**ARL**	**4.52**	**4.27**	**4.10**	**3.99**	**3.93**	**3.87**	**3.83**	**3.95**	**4.41**	**4.61**
	MDRL	4.00	4.00	4.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
	SDRL	2.15	2.16	2.25	2.30	2.39	2.42	2.72	3.10	3.79	4.03
1.7	**ARL**	**3.91**	**3.67**	**3.50**	**3.41**	**3.31**	**3.27**	**3.15**	**3.19**	**3.43**	**3.69**
	MDRL	4.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	3.00	3.00
	SDRL	1.74	1.74	1.81	1.83	1.89	1.95	2.15	2.40	2.78	3.16
1.8	**ARL**	**3.47**	**3.26**	**3.11**	**3.01**	**2.91**	**2.85**	**2.68**	**2.70**	**2.86**	**3.01**
	MDRL	3.00	3.00	3.00	3.00	3.00	2.00	2.00	2.00	2.00	2.00
	SDRL	1.46	1.52	1.52	1.54	1.60	1.63	1.78	1.90	2.26	2.48
1.9	**ARL**	**3.17**	**2.96**	**2.83**	**2.71**	**2.62**	**2.51**	**2.37**	**2.37**	**2.46**	**2.58**
	MDRL	3.00	3.00	3.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
	SDRL	1.30	1.30	1.34	1.39	1.40	1.38	1.47	1.62	1.84	2.00
2.0	**ARL**	**2.92**	**2.73**	**2.59**	**2.43**	**2.37**	**2.29**	**2.13**	**2.12**	**2.16**	**2.24**
	MDRL	3.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
	SDRL	1.13	1.17	1.18	1.18	1.19	1.21	1.29	1.41	1.54	1.69

towards ARL of NEWMA chart with an increase in λ. At $\lambda = 1$, $UCL_t = UCL_a$ as the factor $(1 - (1 - \lambda)^{2t})$ reduces to 1 and hence the ARL performance of both the charts is similar. Moreover Fig. 35.2 shows percentage decrease in ARL_1 of TNEWMA chart as compared to NEWMA chart for certain choices of λ and δ. We can see that the difference in ARL_1 of both the charts is bigger for smaller values of λ and higher values of δ. The difference tends to reduce as λ increases and δ decreases. Hence the use of exact control limits also improves variability EWMA chart performance for detecting shifts of higher magnitude.

Table 35.2 Run length characteristics of TNEWMA chart when $ARL_0 = 200$

λ	L_t	λ									
		0.05	0.10	0.15	0.20	0.25	0.30	0.50	0.70	0.90	1.00
		1.649	1.975	2.164	2.279	2.379	2.440	2.588	2.652	2.685	2.693
1.0	**ARL**	**199.85**	**200.82**	**200.30**	**200.19**	**200.29**	**200.37**	**200.13**	**200.63**	**199.76**	**199.73**
	MDRL	124.00	134.00	136.50	134.00	138.00	137.00	139.00	141.00	140.00	140.00
	SDRL	209.87	206.10	212.73	199.28	208.94	202.42	200.30	199.74	200.94	194.83
1.1	**ARL**	**25.80**	**31.45**	**34.71**	**37.99**	**40.25**	**42.42**	**49.21**	**54.01**	**61.15**	**65.18**
	MDRL	16.00	22.00	25.00	26.00	28.00	30.00	34.00	38.00	43.00	45.00
	SDRL	28.83	32.55	34.31	37.60	40.52	41.58	49.17	53.70	60.63	63.66
1.2	**ARL**	**9.93**	**11.91**	**13.18**	**14.08**	**15.52**	**16.65**	**19.29**	**21.96**	**25.67**	**28.43**
	MDRL	6.00	8.00	10.00	10.00	11.00	12.00	14.00	15.00	18.00	20.00
	SDRL	10.36	11.59	12.53	13.07	14.56	15.26	18.48	21.30	25.29	28.12
1.3	**ARL**	**5.61**	**6.64**	**7.37**	**7.81**	**8.33**	**8.66**	**9.98**	**11.39**	**13.66**	**14.94**
	MDRL	4.00	5.00	6.00	6.00	6.00	6.00	7.00	8.00	10.00	11.00
	SDRL	5.64	6.09	6.54	6.88	7.30	7.73	9.41	10.90	13.12	14.40
1.4	**ARL**	**3.78**	**4.56**	**4.92**	**5.19**	**5.46**	**5.67**	**6.48**	**7.09**	**8.31**	**9.23**
	MDRL	3.00	3.00	4.00	4.00	4.00	4.00	5.00	5.00	6.00	7.00
	SDRL	3.49	3.97	4.21	4.33	4.60	4.69	5.67	6.26	8.05	8.67
1.5	**ARL**	**2.84**	**3.33**	**3.68**	**3.86**	**4.03**	**4.14**	**4.61**	**5.01**	**5.78**	**6.38**
	MDRL	2.00	2.00	3.00	3.00	3.00	3.00	4.00	4.00	4.00	5.00
	SDRL	2.49	2.78	3.01	3.10	3.26	3.28	3.80	4.40	5.24	5.82
1.6	**ARL**	**2.32**	**2.70**	**2.94**	**3.04**	**3.17**	**3.27**	**3.60**	**3.82**	**4.30**	**4.58**
	MDRL	2.00	2.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00
	SDRL	1.90	2.16	2.25	2.30	2.37	2.48	2.82	3.16	3.69	4.17
1.7	**ARL**	**2.00**	**2.29**	**2.43**	**2.55**	**2.66**	**2.68**	**2.92**	**3.07**	**3.38**	**3.65**
	MDRL	1.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	3.00	3.00
	SDRL	1.50	1.70	1.81	1.85	1.95	1.95	2.16	2.41	2.77	3.08
1.8	**ARL**	**1.77**	**2.00**	**2.14**	**2.23**	**2.28**	**2.26**	**2.47**	**2.59**	**2.81**	**3.01**
	MDRL	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
	SDRL	1.23	1.42	1.53	1.58	1.60	1.64	1.78	1.92	2.24	2.47
1.9	**ARL**	**1.60**	**1.78**	**1.91**	**1.97**	**2.02**	**2.03**	**2.18**	**2.27**	**2.39**	**2.56**
	MDRL	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
	SDRL	1.04	1.17	1.28	1.30	1.34	1.35	1.47	1.60	1.90	2.06
2.0	**ARL**	**1.47**	**1.64**	**1.72**	**1.76**	**1.82**	**1.84**	**1.95**	**2.02**	**2.09**	**2.24**
	MDRL	1.00	1.00	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00
	SDRL	0.91	1.03	1.09	1.14	1.17	1.17	1.26	1.33	1.52	1.64

35.4 Effect of Fast Initial Response on Variability EWMA Chart

We have seen in the previous section that the use of time varying control limits as compared to asymptotic limits significantly improves the out-of-control run length behavior of variability EWMA charts. A further increase in the sensitivity of EWMA chart to detect shifts in variability can be achieved by using an FIR feature. The FIR feature, introduced by [29] for CUSUM charts, detects

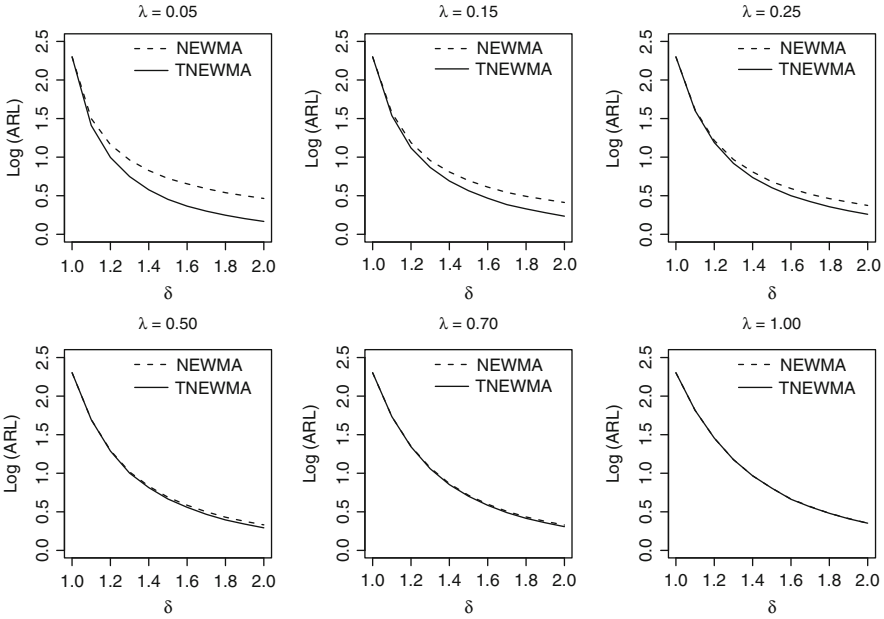


Fig. 35.1 ARL comparison of NEWMA and TNEWMA charts for different values of λ when $ARL_0 = 200$

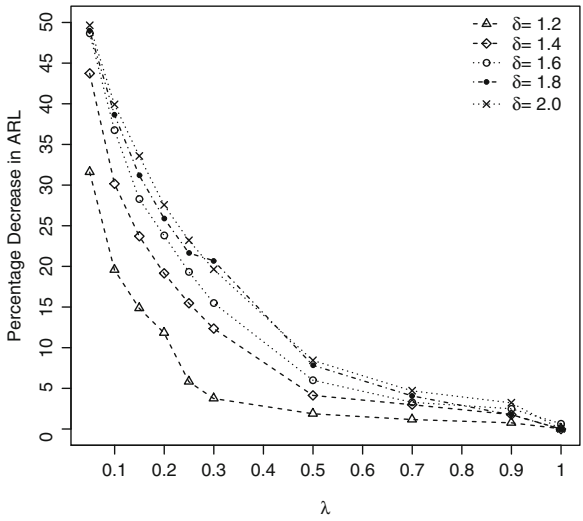


Fig. 35.2 Percentage decrease in out-of-control ARL of TNEWMA chart as compared to NEWMA chart when $ARL_0 = 200$

out-of-control signals more quickly at process startup by assigning some non-zero constant to the starting values of CUSUM chart statistics. Lucas and Saccucci [9] proposed the idea of applying the FIR feature to EWMA control structures by using two one-sided EWMA charts. Rhoads et al. [30] used the

FIR approach for time varying control limits and showed superior performance of their proposed scheme compared to the [9] FIR scheme. Both these schemes were criticized as they require the use of two EWMA charts instead of one for monitoring changes in process parameters. Steiner [11] presented another FIR scheme for EWMA charts. His proposal is based on further narrowing the time varying control limits by using an exponentially decreasing FIR adjustment which is defined as

$$FIR_{adj} = 1 - (1 - f)^{1+a(t-1)}, \tag{35.13}$$

where a is known as the adjustment parameter and is chosen such that the FIR adjustment has very little effect after a specified time period say at $t = 20$, we have $FIR_{adj} = 0.99$. The effect of this FIR adjustment decreases with time and makes the control limit a proportion f of the distance from the starting value [11]. By comparing run length characteristics, Steiner [11] showed that his proposed FIR scheme outperformed the previous FIR schemes by [9, 30]. The FIR adjustment used by [11] is very attractive and has also been recently applied by [31] to generally weighted moving average control charts. In this section we examine the effect of FIR on the performance of variability EWMA chart. The time varying variability EWMA chart using FIR will be referred as the FNEWMA chart for the rest of study. The FNEWMA chart signals an out-of-control condition whenever W_t exceeds UCL_f, where UCL_f is given as

$$UCL_f = L_f\left(1 - (1 - f)^{1+a(t-1)}\right)\sqrt{\frac{\lambda[1 - (1 - \lambda)^{2t}]}{2 - \lambda}}\sigma_{Z_t^+}. \tag{35.14}$$

To obtain a substantial benefit from FIR feature, f should be fairly small. In this study we used $f = 0.5$ and limited the effect of FIR adjustment till $t = 20$ following [11, 31].

The run length characteristics of FNEWMA chart are reported in Table 35.3. ARL_0 for FNEWMA chart is also fixed at 200 by using appropriate L_f values for different choices of λ. By comparing results in Tables 35.1, 35.2 and 35.3, we can observe the superior run length performance of the FNEWMA chart as compared to the NEWMA and TNEWMA charts. For example, the FNEWMA chart has $ARL_1 = 10.72$ for $\lambda = 0.3$ and $\delta = 1.2$, while the corresponding ARL_1 for the TNEWMA and NEWMA charts are 16.65 and 17.30 respectively. This indicates that the FNEWMA chart requires on average nearly six less observations as compared to the NEWMA and TNEWMA charts to detect a shift of 1.2σ in process variability when $\lambda = 0.3$. Figure 35.3 compares the ARLs of the NEWMA, TNEWMA and FNEWMA charts for some choices of λ. We can easily observe that the ARL_1 of the FNEWMA chart is consistently lower than the ARL_1 of both NEWMA and TNEWMA charts for every choice of λ. This indicates that the FNEWMA chart detects shifts in process variability more quickly than the other two charts, the difference seems greater for higher values of λ which is consistent with the findings of [11].

Table 35.3 Run Length Characteristics of FNEWMA chart when $ARL_0 = 200$

δ	L_f	λ									
		0.05	0.10	0.15	0.20	0.25	0.30	0.50	0.70	0.90	1.00
		1.740	2.071	2.241	2.369	2.460	2.530	2.670	2.736	2.770	2.784
1.0	**ARL**	**199.24**	**200.49**	**199.63**	**199.99**	**200.21**	**200.74**	**199.96**	**200.29**	**199.55**	**199.61**
	MDRL	94.00	115.50	117.00	121.00	122.00	122.00	117.50	114.00	109.00	109.00
	SDRL	263.25	249.11	244.59	239.14	240.99	242.92	241.41	249.86	253.24	251.52
1.1	**ARL**	**21.17**	**25.84**	**28.27**	**31.12**	**31.40**	**34.36**	**38.57**	**41.54**	**45.85**	**50.85**
	MDRL	7.00	12.00	14.00	15.00	14.00	15.00	16.00	14.50	14.00	16.00
	SDRL	29.33	34.37	36.66	40.77	41.93	46.10	52.51	58.65	68.32	74.19
1.2	**ARL**	**8.02**	**8.85**	**9.16**	**9.65**	**10.29**	**10.72**	**11.91**	**13.07**	**14.25**	**15.19**
	MDRL	4.00	4.00	4.00	4.00	4.00	5.00	5.00	4.00	4.00	4.00
	SDRL	10.14	11.57	12.57	12.72	13.77	14.46	17.20	20.49	22.95	27.47
1.3	**ARL**	**4.08**	**4.65**	**4.77**	**5.06**	**5.24**	**5.29**	**5.61**	**5.88**	**6.53**	**7.07**
	MDRL	2.00	2.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00
	SDRL	5.15	5.75	5.83	6.16	6.60	6.68	7.53	8.35	10.09	11.57
1.4	**ARL**	**2.76**	**3.12**	**3.28**	**3.28**	**3.29**	**3.39**	**3.45**	**3.56**	**3.78**	**3.96**
	MDRL	1.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
	SDRL	3.18	3.48	3.60	3.65	3.62	3.84	4.07	4.31	5.23	5.63
1.5	**ARL**	**2.11**	**2.33**	**2.42**	**2.48**	**2.50**	**2.54**	**2.58**	**2.51**	**2.65**	**2.80**
	MDRL	1.00	1.00	1.00	1.00	1.00	2.00	2.00	2.00	2.00	2.00
	SDRL	2.07	2.30	2.38	2.49	2.42	2.51	2.60	2.56	2.95	3.34
1.6	**ARL**	**1.80**	**1.94**	**1.99**	**2.02**	**2.06**	**2.06**	**2.09**	**2.12**	**2.13**	**2.19**
	MDRL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SDRL	1.57	1.73	1.74	1.79	1.84	1.83	1.82	1.90	1.93	2.16
1.7	**ARL**	**1.58**	**1.68**	**1.70**	**1.75**	**1.76**	**1.80**	**1.77**	**1.79**	**1.83**	**1.87**
	MDRL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SDRL	1.20	1.30	1.32	1.38	1.38	1.44	1.40	1.38	1.48	1.57
1.8	**ARL**	**1.44**	**1.52**	**1.55**	**1.57**	**1.59**	**1.58**	**1.61**	**1.61**	**1.62**	**1.61**
	MDRL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SDRL	0.96	1.07	1.10	1.12	1.13	1.11	1.12	1.12	1.18	1.16
1.9	**ARL**	**1.34**	**1.40**	**1.43**	**1.44**	**1.47**	**1.46**	**1.46**	**1.48**	**1.49**	**1.50**
	MDRL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SDRL	0.79	0.89	0.91	0.91	0.95	0.93	0.90	0.93	0.93	0.98
2.0	**ARL**	**1.26**	**1.32**	**1.34**	**1.36**	**1.36**	**1.38**	**1.40**	**1.39**	**1.40**	**1.41**
	MDRL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SDRL	0.66	0.74	0.79	0.78	0.77	0.81	0.81	0.80	0.83	0.84

To get more insight into the run length distributions of the NEWMA, TNEWMA and FNEWMA charts, Fig. 35.4 presents run length curves (RLCs) of these charts for certain values of λ using $\delta = 1.2$. We can observe that for smaller values of λ, RLCs of TNEWMA chart are higher than RLCs of NEWMA chart indicating that TNEWMA chart has greater probability for shorter run lengths for these λ values. The superiority of FNEWMA chart over NEWMA and TNEWMA charts is also clear for all values of λ. Note that this high probability at shorter run lengths indicate that the shifts in the process variability will be detected quickly with high probability.

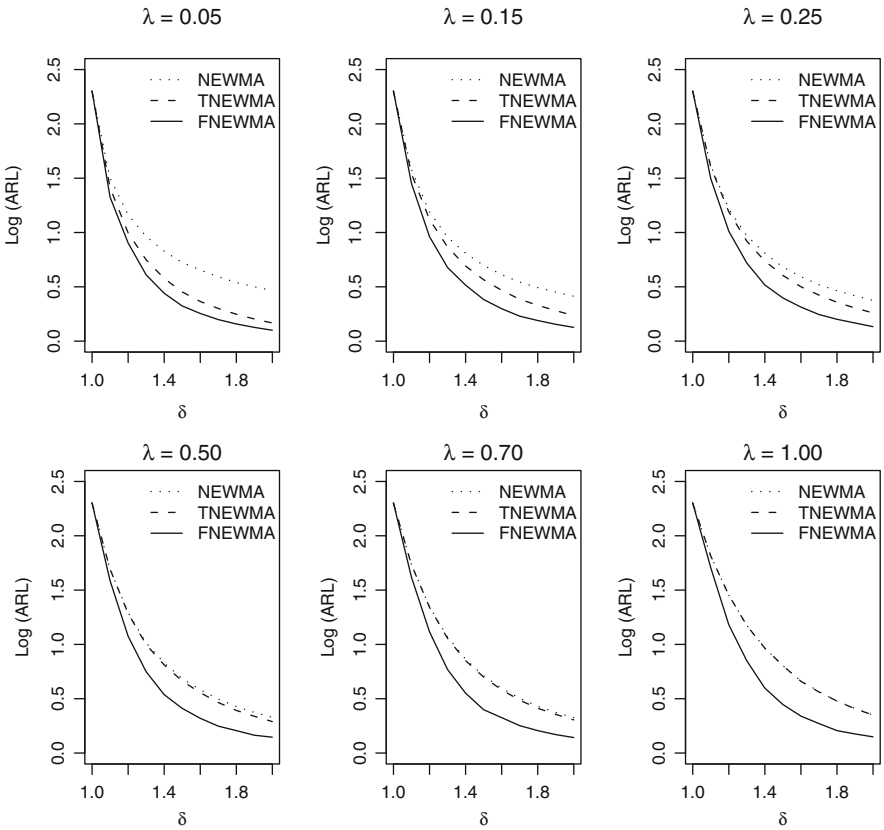


Fig. 35.3 ARL comparison of NEWMA, TNEWMA and FNEWMA charts for different values of λ when $ARL_0 = 200$

35.5 Conclusions

This chapter examines the performance of variability EWMA chart using asymptotic, time varying and FIR based control limits. It has been shown that the ability of the variability EWMA chart to detect shifts in variation can be improved by using exact (time varying limits) instead of asymptotic control limits, particularly for smaller values of smoothing parameter λ. The FIR feature has also shown to contribute significantly to further increase the sensitivity of the EWMA chart to detect shifts in process variability. Computations have been performed using NEWMA chart but these results can be generalized for the other variability EWMA charts discussed in Sect. 35.1. This study will help quality practitioners to choose a more sensitive variability EWMA chart.

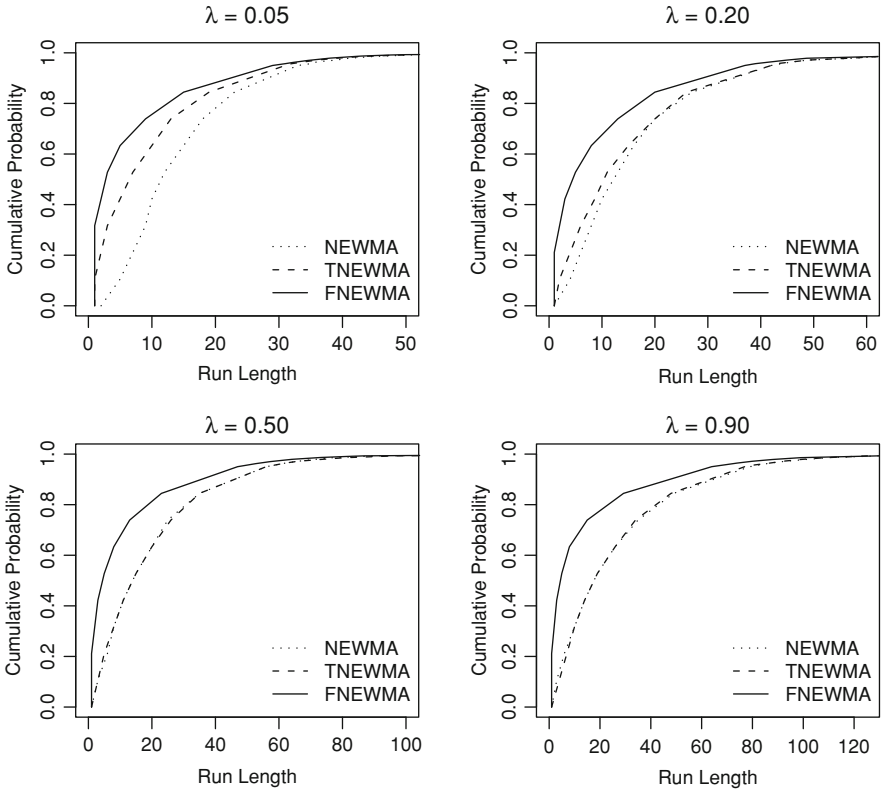


Fig. 35.4 Run length curves of NEWMA, TNEWMA and FNEWMA charts for different values of λ when $\delta = 1.2$ and $ARL_0 = 200$

References

1. Hwang SL, Lin JT, Liang GF, Yau YJ, Yenn TC, Hsu CC (2008) Application control chart concepts of designing a pre-alarm system in the nuclear power plant control room. Nucl Eng Design 238(12):3522–3527
2. Woodall WH (2006) The use of control charts in health-care and public-health surveillance. J Qual Technol 38(2):89–104
3. Wang Z, Liang R (2008) Discuss on applying SPC to quality management in university education. In: Proceedings of the 9th international conference for young computer scientists, ICYCS 2008, pp 2372–2375
4. Masson P (2007) Quality control techniques for routine analysis with liquid chromatography in laboratories. J Chromatogr A 1158(1–2):168–173
5. Abbasi SA (2010) On the performance of EWMA chart in presence of two component measurement error. Qual Eng 22(3):199–213
6. Montgomery DC (2001) Introduction to statistical quality control, 4th edn. Wiley, New York
7. Ryan PR (2000) Statistical methods for quality improvement, 2nd edn. Wiley, New York
8. Roberts SW (1959) Control chart tests based on geometric moving averages. Technometrics 1(3):239–250

9. Lucas JM, Saccucci MS (1990) Exponentially weighted moving average control schemes. Properties and enhancements. Technometrics 32(1):1–12
10. Montgomery DC, Torng JCC, Cochran JK, Lawrence FP (1995) Statistically constrained economic design of the EWMA control chart. J Qual Technol 27(3):250–256
11. Steiner SH (1999) EWMA control charts with time-varying control limits and fast initial response. J Qual Technol 31(1):75–86
12. Chan LK, Zhang J (2000) Some issues in the design of EWMA charts. Commun Stat Part B Simul Comput 29(1):207–217
13. Maravelakis PE, Panaretos J, Psarakis S (2004) EWMA chart and measurement error. J Appl Stat 31(4):445–455
14. Carson PK, Yeh AB (2008) Exponentially weighted moving average (EWMA) control charts for monitoring an analytical process. Ind Eng Chem Res 47(2):405–411
15. Shu L, Jiang W (2008) A new EWMA chart for monitoring process dispersion. J Qual Technol 40(3):319–331
16. Abbasi SA (2010) On sensitivity of EWMA control chart for monitoring process dispersion. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, vol III, WCE 2010, 30 June–2 July, 2010, London, UK, pp 2027–2032
17. Wortham AW, Ringer LJ (1971) Control via exponential smoothing. Transportation Logistic Rev 7:33–39
18. Domangue R, Patch SC (1991) Some omnibus exponentially weighted moving average statistical process monitoring schemes. Technometrics 33:299–313
19. Crowder SV, Hamilton M (1992) Average run lengths of EWMA controls for monitoring a process standard deviation. J Qual Technol 24:44–50
20. MacGregor JF, Harris TJ (1993) The exponentially weighted moving variance. J Qual Technol 25:106–118
21. Stoumbos ZG, Reynolds MR Jr (2000) Robustness to non normality and autocorrelation of individual control charts. J Stat Comput Simul 66:145–187
22. Chen GM, Cheng SW, Xie HS (2001) Monitoring process mean and variability with one EWMA chart. J Qual Technol 33:223–233
23. Barr DR, Sherrill ET (1999) Mean and variance of truncated normal distributions. Am Stat 53:357–361
24. Maravelakis P, Panaretos J, Psarakis S (2005) An examination of the robustness to non-normality of the EWMA control charts for the dispersion. Commun Stat Simul Comput 34(4):1069–1079
25. Neubauer AS (1997) The EWMA control chart: properties and comparison with other quality-control procedures by computer simulation. Clin Chem 43(4):594–601
26. Zhang L, Chen G (2004) EWMA charts for monitoring the mean of censored Weibull lifetimes. J Qual Technol 36(3):321–328
27. Kim MJ (2005) Number of replications required in control chart Monte Carlo simulation studies. PhD Dissertation, University of Northern Colorado
28. Schaffer JR, Kim MJ (2007) Number of replications required in control chart Monte Carlo simulation studies. Commun Stat Simul Comput 36(5):1075–1087
29. Lucas JM, Crosier RB (1982) Fast initial response for CUSUM quality control schemes: give your CUSUM a head start. Technometrics 24(3):199–205
30. Rhoads TR, Montgomery DC, Mastrangelo CM (1996) A fast initial response scheme for the exponentially weighted moving average control chart. Qual Eng 9(2):317–327
31. Chiu WC (2009) Generally weighted moving average control charts with fast initial response features. J Appl Stat 36(3):255–275

Chapter 36
Assessing Response's Bias, Quality of Predictions, and Robustness in Multiresponse Problems

Nuno Costa, Zulema Lopes Pereira and Martín Tanco

Abstract Optimization measures for evaluating compromise solutions in multiresponse problems formulated in the Response Surface Methodology framework are proposed. The measures take into account the desired properties of responses at optimal variable settings, namely, the bias, quality of predictions and robustness, which allow the analyst to achieve compromise solutions of interest and feasible in practice, namely in the case of a method that does not consider in the objective function the responses' variance level and correlation information is used. Two examples from the literature show the utility of the proposed measures.

36.1 Introduction

Statistical tools and methodologies like the response surface methodology (RSM) have been increasingly used in industry and became a change agent in the way design and process engineers think and work [9]. In particular, RSM has been used for developing more robust systems (process and product), improving and

N. Costa (✉)
Setúbal Polytechnic Institute, College of Technology, Campus do IPS, Estefanilha, 2910-761 Setúbal, Portugal
e-mail: nuno.costa@estsetubal.ips.pt

N. Costa · Z. L. Pereira
UNIDEMI/DEMI, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
e-mail: zlp@fct.unl.pt

M. Tanco
CITEM, Universidad de Montevideo, Luis P. Ponce, 1307, 11300 Montevideo, Uruguay
e-mail: mtanco@um.edu.uy

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_36,
© Springer Science+Business Media B.V. 2011

optimizing systems performance with the required efficiency and effectiveness. The readers are referred to Myers et al. [18] for a thoroughly discussion on this methodology.

While most case studies reported in the literature focus on the optimization of one single quality characteristic of process or product, the variety of real-life problems requires the consideration of multiple quality characteristics (objectives; responses). This fact and the researchers' desire to propose enhanced techniques using recent advancements in mathematical optimization, scientific computing and computer technology have been making the multiresponse optimization an active research field. New algorithms and methodologies have been developed and their diffusion into various disciplines has proceeded at a rapid pace. To date, researchers are paying great attention to hybrid approaches to avoid premature algorithm convergence toward a local maximum or minimum and reach the global optimum in problems with multiple responses [24]. The readers are referred to Younis and Dong [25] for a review on historical development, special features and trends on the development of global optimization algorithms. These authors also examine and compare a number of representatives and recently introduced global optimization techniques. The issue is that the level of computational and mathematical or statistical expertise required for using those algorithms or methodologies and solving such problems successfully is significant. This makes such sophisticated tools hard to adopt, in particular, by practitioners [1].

A strategy widely used for optimizing multiple responses in the RSM framework consists of converting the multiple responses into a single (composite) function followed by its optimization, using either the generalized reduced gradient or sequential quadratic programming algorithms available in the popular Microsoft Excel® (Solver add-in) and Matlab® (fmincon routine), respectively. To form that composite function, the desirability function-based and loss function-based methods are the most popular among practitioners.

The existing methods use distinct composite functions to provide indication about how close the response values are from their target, but the widely used desirability-based methods do not consider the responses' variance level and correlation information, and the composite function does not give information on it to the analyst. What the analyst knows is that either a higher or a lower value is preferred, depending on how the composite function is defined. The composite functions of the loss function-based methods present the result in monetary terms, that is, the compromise solution is expressed by a monetary loss that must be as low as possible, and some of those composite functions consider the variance–covariance structure of responses. However, the composite functions of loss- and desirability-based methods have a serious drawback. They may give inconsistent results, namely, different results for the same responses values (compromise solution). This may confound the analyst and difficult the evaluation of compromise solutions as he/she needs to check the values of each response considered in the study to identify a compromise solution of interest. This difficulty increases as larger the number of available solutions and responses are. So, the authors propose optimization performance measures with a threefold purpose:

I. Provide relevant information to the analyst so that he/she may achieve compromise solutions of interest and feasible in practice whenever a method that does not consider in the objective function the variance–covariance structure of responses is used.
II. Help the analyst in evaluating the feasibility of compromise solutions by assessing the response's bias (responses deviation from their target), quality of predictions (variance due to uncertainty in the regression coefficients of predicted responses) and robustness (variance due to uncontrollable variables) separately.
III. Allow the evaluation of methods solutions that cannot be compared directly due to the different approaches subjacent to those methods, for example, loss function and desirability function approaches.

The feasibility of the proposed measures is illustrated through desirability and loss function-based methods by using two examples from the literature.

The remaining sections are structured as follows: next section provides a review on analysis methods. Then optimization measures are introduced. The subsequent sections include the examples and the results discussion, respectively. Conclusion and future work are presented in the last section.

36.2 Methods for Multiresponse Analysis

The desirability function-based and loss function-based methods are the most popular ones among practitioners who, in the RSM framework, look for optimum variable settings for the process and product whenever multiple responses are considered simultaneously. Therefore, the methods that are widely used in practice and will serve to illustrate the feasibility of optimization measures are reviewed below. Many other alternative approaches are available in the literature, and reviews on them are provided by Gauri and Pal [8], Kazemzadeh et al. [10], Murphy et al. [16].

36.2.1 Desirability-Based Methods

The desirability-based methods are easy to understand, flexible for incorporating the decision-maker's preferences (priority to responses), and the most popular of them, the so-called Derringer and Suich's method [6], or modifications of it [5], is available in many data analysis software packages. However, to use this method the analyst needs to assign values to four shape parameters (weights). This is not a simple task and makes an impact on the optimal variable settings. An alternative desirability-based method that, under the assumptions of normality and homogeneity of error variances, requires minimum information from the user was

proposed by Ch'ng et al. [2]. The method they proposed is easy to understand and implement in the readily available Microsoft Excel®—Solver tool and, in addition, requires less cognitive effort from the analyst. The user only has to assign values to one type of shape parameters (weights), which is a relevant advantage over the extensively used Derringer and Suich's method.

Ch'ng et al. [2] suggested individual desirability functions of the form

$$d = \frac{2\hat{y} - (U + L)}{U - L} + 1 = \frac{2\hat{y}}{U - L} + \frac{-2L}{U - L} = m\hat{y} + c \tag{36.1}$$

where $0 \leq d \leq 2$ and $\hat{y}$ represents the response's model with upper and lower bounds defined by U and L, respectively. The global desirability (composite) function is defined as

$$D = \left(\sum_{i=1}^{p} e_i |d_i - d_i(\theta_i)| \right) \Big/ p \tag{36.2}$$

where $d_i(\theta_i)$ is the value of the individual desirability function i at the target value θ_i, e_i is the weight (degree of importance or priority) assigned to response i, p is the number of responses, and $\sum_{i=1}^{p} e_i = 1$. The aim is to minimize D.

Although Ch'ng et al. illustrate their method only for nominal-the-best (NTB—the value of the estimated response is expected to achieve a particular target value) response type, in this article the larger-the-best (LTB—the value of the estimated response is expected to be larger than a lower bound) and smaller-the-best (STB—the value of the estimated response is expected to be smaller than an upper bound) response types are also considered. In these cases, $d_i(U_i)$ and $d_i(L_i)$ are used in Eq. 36.2 instead of $d_i(\theta_i)$, under the assumption that it is possible to establish the specification limits U and L to those responses based on product knowledge or practical experience. To use the maximum or minimum value of the response model is also an alternative. A limitation in this method is that it does not consider the quality of predictions and robustness in the optimization process.

36.2.2 Loss Function-Based Methods

The loss function approach uses a totally different idea about the multi-response optimization by considering monetary aspects in the optimization process. This approach is very popular among the industrial engineering community and, unlike the above-mentioned desirability-based methods, there are loss function-based methods that consider the responses' variance level and exploit the responses' correlation information, which is statistically sound. Examples of those methods were introduced by Vining [21] and Lee and Kim [13].

Vining [21] proposed a loss function-based method that allows specifying the directions of economic importance for the compromise optimum, while seriously

considering the variance–covariance structure of the expected responses. This method aims at finding the variable settings that minimize an expected loss function defined as

$$E[L(\hat{y}(x),\theta)] = (E[\hat{y}(x)] - \theta)^T C\,(E[\hat{y}(x)] - \theta) + \text{trace}\left[C \sum_{\hat{y}}(x)\right] \tag{36.3}$$

where $\sum_{\hat{y}}(x)$is the variance–covariance matrix of the predicted responses at x and C is a cost matrix related to the costs of non-optimal design. If C is a diagonal matrix then each element represents the relative importance assigned to the corresponding response, that is, the penalty (cost) incurred for each unit of response value deviated from its optimum. If C is a non-diagonal matrix, the off-diagonal elements represent additional costs incurred when pairs of responses are simultaneously off-target. The first term in Eq. 36.3 represents the penalty due to the deviation from the target; the second term represents the penalty due to the quality of predictions.

Lee and Kim [13], such as Pignatiello [19] and Wu and Chyu [22], emphasize the bias reduction and the robustness improvement. They proposed minimizing an expected loss defined as

$$E[L(y(x),\theta)] = \sum_{i=1}^{p} c_i\left[(\hat{y}_i - \theta_i)^2 + \hat{\sigma}_i^2\right] + \sum_{i=2}^{p}\sum_{j=1}^{i-1} c_{ij}\left[\hat{\sigma}_{ij} + (\hat{y}_i - \theta_i)(\hat{y}_j - \theta_j)\right] \tag{36.4}$$

where c_i and c_{ij} represent weights (priorities or costs), and $\hat{\sigma}_i^2$ and $\hat{\sigma}_{ij}$ are elements of the response's variance–covariance structure at x ($\sum_y(x)$). A key difference between Eqs. 36.3 and 36.4 is that the later uses the variance–covariance structure of the responses rather than the variance–covariance structure of the predicted responses. Moreover, Lee and Kim's method requires replicates at each design run, which will certainly increase the time and cost of experimentation. This is not problematic only if the variance due to uncontrollable variables is a trouble in practice.

A difficulty with all loss function-based methods is to take into account different scales, relative variabilities and relative costs in matrix C [12, 21].

36.3 Measures of Optimization Performance

To evaluate the feasibility of compromise solutions in multiresponse problems, the analyst needs information about the response's properties at "optimal" variable settings, namely the bias and variance. In fact, responses at some variable settings may have considerable variance due to the uncertainty in the regression coefficients of predicted responses and are sensitive to uncontrollable variables that may be significant and, therefore, cannot be ignored.

In the RSM framework few authors have addressed the evaluation of response's properties to the extent it deserves. In general they focus on the output of the objective function they use. Authors that compare the performance of several methods by evaluating the responses properties at variable settings through optimization performance measures are Lee and Kim [13], Ko et al. [12] and Xu et al. [23]. While Lee and Kim [13] and Ko et al. [12] use the terms or components of the objective function they propose for comparing the results of loss function-based methods in terms of the desired response's properties, Xu et al. [23] propose new optimization performance measures. The major shortcomings in the proposals of previous authors are the following:

1. The optimization measures used by Lee and Kim [13] and Ko et al. [12] require the definition of a cost matrix, which is not easy to define or readily available.
2. The optimization measures used by Xu et al. [23] only allow the evaluation of response's bias.

To compare methods results or compromise solutions in multiresponse optimization problems it is necessary to consider the statistical properties of the methods used in addition to response's bias and variance. In fact, optimization methods may differ in terms of statistical properties and optimization schemes so the evaluation and comparison of the corresponding solutions in a straightforward manner may not be possible. For example, the global desirability values of methods that either minimize or minimize the global desirability are neither comparable directly nor with the result (monetary loss) achieved from a loss function-based method.

With the aim at providing useful information to the analyst or decision-maker concerning to desired response's properties (bias, quality of predictions, and robustness) and to evaluate the solutions obtained from different methods optimization measures are proposed. Those measures allow the separate assessment of the bias, quality of predictions, and robustness, which may help the analyst in achieving a solution of interest and guiding him/her during the optimization process, in particular when quality of predictions and robustness are important issues in practice. In addition, they may also serve to evaluate the solutions obtained from different methods and help the practitioner or researcher in making a more informed decision when he/she is interested in choosing a method for optimizing multiple responses.

To assess the method's solutions in terms of bias, it is suggested an optimization measure that considers the response types, response's specification limits and deviation of responses from their target. This measure, named cumulative bias (B_{cum}), is defined as

$$B_{cum} = \sum_{i=1}^{p} W_i \left| \hat{y}_i^* - \theta_i \right| \tag{36.5}$$

where $\hat{y}_i^*$ represents the estimated response value at "optimal" variable settings, θ_i is the target value and W_i is a parameter that takes into account the

specification limits and response type of the *i*th response. This parameter is defined as follows: $W = 1/(U - L)$ for STB and LTB response types; $W = 2/(U - L)$ for NTB response type.

The cumulative bias gives an overall result of the optimization process instead of focusing on the value of a single response, which prevents unreasonable decisions of being taken in some cases [11]. To assess the bias of each response, the practitioner may use the individual bias (B_i) defined as

$$B_i = W_i \left| \hat{y}_i^* - \theta_i \right| \tag{36.6}$$

Alternatives to B_{cum} and B_i are presented by Xu et al. [23]. These authors utilize $W_i = 1/\theta_i$ and consider the mean value for the cumulative bias. For the individual bias they consider $W_i = 1$.

The measure proposed for assessing method's solutions in terms of quality of predictions is defined by

$$\text{QoP} = \text{trace}\left[\varphi \sum\nolimits_{\hat{y}} (x)\right] = \text{trace}\left[\varphi\left(x_j^T \left[X^T Q^{-1} X\right]^{-1} x_j\right)\right] \tag{36.7}$$

where x_j is the subset of independent variables consisting of the K × 1 vector of regressors for the *i*th response with N observations on K_i regressors for response, X is an $Np \times K$ block diagonal matrix and $Q = \sum \otimes \mathrm{I}_N$. An estimate of $\sum$ is $\hat{\sigma}_{ij} = \hat{e}_j^T \hat{e}_j / N$, where $\hat{e}$ is the residual vector from the OLS estimation of the response i; I_N is an identity matrix and $\otimes$ represents the Kronecker product. To make $\sum_{\hat{y}} (x)$ dimensionless this matrix is multiplied by matrix φ, whose diagonal and non-diagonal elements are $\varphi_{ii} = 1/(U_i - L_i)^2$ and $\varphi_{ir} = 1/(U_i - L_i)(U_r - L_r)$ for $i \neq r$, respectively.

QoP is defined under the assumption that seemingly unrelated regression (SUR) method is employed to estimate the regression models (response surfaces) as it yields regression coefficients at least as accurate as those of other popular regression techniques, namely the ordinary and generalized least squares [7, 20]. If the ordinary least squares is used the reader is referred to Vining [21] as this author presents variants of Eq. 36.7 for the case of regression models with equal and different forms.

The robustness is assessed by

$$\text{Rob} = \text{trace}\left[\varphi \sum\nolimits_{y} (x)\right] \tag{36.8}$$

where $\sum_y (x)$ represents the variance–covariance matrix of the "true" responses. Note that replications of the experimental runs are required to assess the robustness and matrix (only considers the specification limits of the variance models, while Lee and Kim [13] and Ko et al. [12] use a cost matrix (C). Although the replicates increase the time and cost of experimentation, they may provide significant improvements in robustness that overbalance or at least compensate the time and cost spent.

B_i, B_{cum}, QoP and Rob are dimensionless ratios, so the worry with the dimensional consistency of responses is cancelled. These measures do not exclude others from being used as well and, in terms of results, the lower their values are, the better the compromise solution will be. In practice, all the proposed measures take values greater than or equal to zero, but zero is the most favorable.

36.4 Examples

Two examples from the literature illustrate the utility of the proposed performance measures. The first one considers a case study where the quality of prediction is the adverse condition. In this example the methods introduced by Ch'ng et al. [2] and Vining [21] are used. The second one considers the robustness as adverse condition. In this case the methods introduced by Ch'ng et al. [2] and Lee and Kim [13] are used.

Example 1 The responses specification limits and targets for the percent conversion (y_1) and thermal activity (y_2) of a polymer are the following: $\hat{y}_1 \geq 80.00$ with $U_1 = \theta_1 = 100$; $55.00 \leq \hat{y}_2 \leq 60.00$ with $\theta_2 = 57.50$. Reaction time (x_1), reaction temperature (x_2), and amount of catalyst (x_3) are the control factors. According to Myers and Montgomery [17], the objective was to maximize the percent conversion and achieve the nominal value for the thermal activity. A central composite design with six axial and six center points, with $-1.682 \leq x_i \leq 1.682$, was run to generate the data. The predicted responses, fitted by the SUR method, are as follows:

$$\begin{aligned}\hat{y}_1 = {} & 81.09 + 1.03x_1 + 4.04x_2 + 6.20x_3 - 1.83x_1^2 + 2.94x_2^2 - 5.19x_3^2 \\ & + 2.13x_1x_2 + 11.38x_1x_3 - 3.88x_2x_3\end{aligned}$$

$$\begin{aligned}\hat{y}_2 = {} & 59.85 + 3.58x_1 + 0.25x_2 + 2.23x_3 - 0.83x_1^2 + 0.07x_2^2 - 0.06x_3^2 \\ & - 0.39x_1x_2 - 0.04x_1x_3 + 0.31x_2x_3\end{aligned}$$

The model of the thermal activity includes some insignificant regressors (x_2, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3), so the predicted response has a poor quality of prediction. In particular, this estimated response will have a variance as larger as farther from the origin the variable settings are. The variance–covariance matrix is estimated as

$$\hat{\Sigma} = \begin{bmatrix} 11.12 & -0.55 \\ -0.55 & 1.55 \end{bmatrix}$$

.

As regards the results, Table 36.1 shows that the global desirability function (D) yields different values for the same response values (cases I and III). This is not desirable or reasonable and may confound analysts who are focused on D value for making decisions. In contrast, the B_{cum} and QoP remain unchanged, as it is expectable in these instances. By using these measures the analyst can easily

Table 36.1 Results: Example 1

	Ch'ng et al.			Vining
	Case I	Case II	Case III	
Weights	(0.30, 0.70)	(0.50, 0.50)	(0.60, 0.40)	$\begin{bmatrix} 0.100 & 0.025 \\ 0.025 & 0.500 \end{bmatrix}$
x_i	(−0.544, 1.682, −0.599)	(−1.682, 1.682, −1.059)	(−0.538, 1.682, −0.604)	(−0.355, 1.682, −0.468)
$\hat{y}_i$	(95.19, 57.50)	(98.04, 55.00)	(95.19, 57.50)	(95.24, 58.27)
Result	$D = 0.14$	$D = 0.35$	$D = 0.29$	$E(\text{loss}) = 3.86$
B_{cum}	0.24	1.10	0.24	0.55
B_i	(0.24, 0.00)	(0.10, 1.00)	(0.24, 0.00)	(0.24, 0.31)
QoP	0.08	0.31	0.08	0.06

perceive whether the changes he/she made in the weights are either favorable or unfavorable in terms of response values. When B_{cum} or QoP increase, it means that the changes made in the weights are unfavorable, that is, the value of at least one of the responses is farther from its target, as it is the case of $\hat{y}_2$ in the Vining's solution, or the quality of predictions is lower, such as occur in case II.

Case II serves to illustrate that the analyst can distinguish solutions with larger variability from other(s) with smaller variability, for example the cases I and III, looking at QoP value.

Vining's solution is the best in terms of the QoP value, because x_1 and x_3 values are slightly closer to the origin than in the other cases, namely the cases I and III, which present the same value of QoP.

These results provide evidence that the proposed measures give better indications (information) to the analyst and can help him/her in achieving feasible solutions if the quality of predictions is an adverse condition.

Example 2 Lee and Kim [13] assumed that the fitted response functions for process mean, variance and covariance of two quality characteristics are as follows:

$$\hat{y}_1 = 79.04 + 17.74x_1 + 0.62x_2 + 14.79x_3 - 0.70x_1^2 - 10.95x_2^2 - 0.10x_3^2 - 5.39x_1x_2 + 1.21x_1x_3 - 1.79x_2x_3$$

$$\hat{\sigma}_1 = 4.54 + 3.92x_1 + 4.29x_2 + 1.66x_3 + 1.15x_1^2 + 4.40x_2^2 + 0.94x_3^2 + 3.49x_1x_2 + 0.74x_1x_3 + 1.19x_2x_3$$

$$\hat{y}_2 = 400.15 - 95.21x_1 - 28.98x_2 - 55.99x_3 + 20.11x_1^2 + 26.80x_2^2 + 10.91x_3^2 + 57.13x_1x_2 - 3.73x_1x_3 - 10.87x_2x_3$$

$$\hat{\sigma}_2 = 26.11 - 1.34x_1 + 6.71x_2 + 0.37x_3 + 0.77x_1^2 + 2.99x_2^2 - 0.97x_3^2 - 1.81x_1x_2 + 0.41x_1x_3$$

Table 36.2 Results: Example 2

	Lee and Kim			Ch'ng et al.
	Case I	Case II	Case III	
Weights	(1, 1, 1)	(0.3, 0.5, 0.02)	(0.8, 0.3, 1.0)	(0.25, 0.25, 0.15, 0.35)
x_i	(0.79, −0.76, 1.00)	(0.80, −0.77, 1.00)	(1.00, −1.00, −0.43)	(0.80, −0.75, 1.00)
$\hat{y}_i$	(97.86, 301.40)	(98.06, 300.32)	(74.22, 346.45)	(98.18, 300.00)
Var–cov	(7.80, 22.96, 6.39)	(7.84, 22.98, 6.39)	(5.89, 23.12, 4.35)	(7.86, 22.99, 6.38)
Result	$E(\text{loss}) = 598.1$	$E(\text{loss}) = 283.7$	$E(\text{loss}) = 173.9$	$D = 0.53$
B_{cum}	1.76	1.75	2.39	1.75
B_i	(0.05, 0.78, 0.01, 0.92)	(0.05, 0.78, 0.00, 0.92)	(0.64, 0.59, 0.23, 0.92)	(0.05, 0.79, 0.00, 0.92)
Rob	0.053	0.053	0.036	0.053

$$\hat{\sigma}_{12} = 5.45 - 0.77x_1 + 0.16x_2 + 0.49x_3 - 0.42x_1^2 + 0.50x_2^2 - 0.35x_3^2 - 0.63x_1x_2 + 1.13x_1x_3 - 0.30x_2x_3$$

In this example it is assumed that the response's specifications are: $\hat{y}_1 \geq 60$ with $U_1 = \theta_1 = 100$; $\hat{y}_2 \leq 500$ with $L_2 = \theta_2 = 300$; $\hat{\sigma}_1 \leq 10$ with $L_1 = \theta_1 = 0$; $\hat{\sigma}_2 \leq 25$ with $L_2 = \theta_2 = 0$; $-1 \leq x_i \leq 1$.

As regards the results, Table 36.2 shows that the loss function proposed by Lee and Kim yields different expected loss values for solutions with marginal differences in the response values (cases I and II). In contrast, the B_{cum} value remains unchanged in these situations, confirming its utility for assessing compromise solutions for multiresponse problems. Moreover, note that the lowest expected loss value is obtained from a solution with the worse values for $\hat{y}_1$ and $\hat{y}_2$, such as occurs in case III, what is an absurdity. Nevertheless, this example provides evidence that the analyst can recognize more robust solutions (case III) from others with larger variability due to uncontrollable factors (case I, II, and Ch'ng et al.'s solution) looking at the Rob value. Note that Ch'ng et al.'s method yields a solution similar to the cases I and II when appropriate weights are assigned to $\hat{y}_2$ and $\hat{\sigma}_2$, remaining unchanged (equal to 0.25) the weights to $\hat{y}_1$ and $\hat{\sigma}_1$.

This example confirms that the proposed measures give useful information to the analyst and can help him/her in achieving feasible solutions if the robustness is an adverse condition.

36.5 Discussion

The optima are stochastic by nature and understanding the variability of responses is a critical issue for the practitioners. Thus, the assessment of the responses' sensitivity to uncontrollable factors in addition to estimated responses' variance

level at "optimal" variable settings by appropriate measures provides the required information for the analyst evaluating compromise solutions in multiresponse optimization problems. For this purpose the QoP and Rob measures are introduced, in addition to measures for assessing the response's bias (B_i and B_{cum}).

The previous examples show that the expected loss and global desirability functions may give inconsistent and incomplete information to the analyst about methods solutions, namely in terms of the merit of the final solution and desired responses properties. This is a relevant shortcoming, which is due to the different weights or priorities assigned to responses that are considered in the composite function. Those composite functions yield different results in cases where the solutions are equal or have slightly changes in the response values, such as illustrated in Examples 1and 2.

Example 2 also shows that absurd results may occur in loss functions if the elements of matrix C are not defined properly. In fact, the loss coefficients (c_{ij}) play a major role in the achievement of optimal parameter conditions that result in trade-offs of interest among responses [22]. In particular, the non-diagonal elements represent incremental costs incurred when pairs of responses are simultaneously off-target, and have to satisfy theoretical conditions that the practitioner may not be aware of or take into account. Those conditions for symmetric loss functions are: $c_{11}, c_{22} \geq 0$ and $-2c_{11}c_{22} \leq c_{12} \leq c_{11}c_{22}$. When these conditions are not satisfied, worse solutions may produce spuriously better (lower) values in the loss function, as it was illustrated with case III in Example 2. Wu and Chyu [22] provide guidelines for defining the c_{ij} for symmetric and asymmetric loss functions, but additional subjective information is required from the analyst.

Therefore, if the analyst only focuses on the result of the composite function used for making decisions he/she may ignore a solution of interest or be confounded about the directions for changing weights or priorities to responses as the composite function may give unreliable information. By using the proposed measures the analyst does not have to worry with the reliability of the information as they do not depend on priorities assigned to responses. By this reason, the proposed measures may also serve to compare the performance of methods that use different approaches, for example, between desirability function-based methods and loss function-based methods. Similarly, they make possible the comparison between methods structured under the same approach but that use different composite functions, as it is the case of Derringer and Suich's method, where the composite function is a multiplicative function, which must be maximized, and Ch'ng et al.'s method, where the composite function is an additive function, which must be minimized.

From a theoretical point of view, methods that consider the responses' variance level and exploit the responses' correlation information lead to solutions that are more realistic when the responses have either significantly different variance levels or are highly correlated [12]. However, the previous examples show that the proposed measures can provide useful information to the analyst so that he/she achieves compromise solutions with desired properties at "optimal" settings by using

methods that do not consider in the objective function the variance–covariance structure of responses. Nevertheless, it is important to highlight that points in non-convex response surfaces cannot be captured by weighted sums like those represented by the objective functions reviewed here, even if the proposed measures are used. Publications where this and other method's limitations are addressed include Das and Dennis [4], Mattson and Messac [14]. This means that the proposed optimization measures are not the panacea to achieve optimal solutions. In fact, Messac et al. [15] demonstrated that the ability of an objective function to capture points in convex and concave surfaces depends on the presence of parameters that the analyst can use to manipulate the composite function's curvature. Although they show that using exponents to assign priorities to responses is a more effective practice to capture points in convex and highly concave surfaces, assigning weights to responses is a critical task in multiresponse problems. It usually involves an undefined trial-and-error weight-tweaking process that may be a source of frustration and significant inefficiency, particularly when the number of responses and control factors is large. So, the need for methods where minimum subjective information is required from the analyst is apparent. A possible choice is the method proposed by Costa [3]. According to this author, besides the low number of weights required from the analyst, the method he proposes has three major characteristics: effectiveness, simplicity and application easiness. This makes the method appealing to use in practice and support the development of an iterative procedure to achieve compromise solutions to multiresponse problems in the RSM framework. Despite the potential usefulness of interactive procedures in finding compromise solutions of interest, due attention has not been paid to procedures that facilitate the preference articulation process for multiresponse problems in the RSM framework.

36.6 Conclusion and Future Work

Low bias and minimum variance are desired response's properties at optimal variable settings in multiresponse problems. Thus, optimization performance measures that can be utilized with the existing methods to facilitate the evaluation of compromise solutions in terms of the desired response's properties are proposed. They can be easily implemented by the analysts and allow the separate assessment of the bias, quality of predictions, and robustness of those solutions. This is useful as the analyst can explore the method's results putting emphasis on the property(ies) of interest. In fact, compromise solutions where some responses are more favorable than others in terms of bias, quality of predictions or robustness may exist. In these instances, the analyst has relevant information available to assign priorities to responses and make a more informed decision based on economical and technical considerations. As the assignment of priorities to responses is an open research field, an iterative procedure that considers the results of the proposed optimization measures arises as an interesting research topic.

References

1. Ayvaz M, Tamer K, Ali H, Ceylan H, Gurarslan G (2009) Hybridizing the harmony search algorithm with a spreadsheet 'Solver' for solving continuous engineering optimization problems. Eng Optim 41(12):1119–1144
2. Ch'ng C, Quah S, Low H (2005) A new approach for multiple response optimization. Qual Eng 17(4):621–626
3. Costa N (2010) Simultaneous optimization of mean and standard deviation. Qual Eng 22(3):140–149
4. Das I, Dennis J (1997) A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. Struct Optim 14(1):63–69
5. Derringer G (1994) A balancing act: optimizing product's properties. Qual Prog 24:51–58
6. Derringer G, Suich R (1980) Simultaneous optimization of several response variables, J Qual Tech 12(4):214–218
7. Fogliatto F, Albin L (2000) Variance of predicted response as an optimization criterion in multiresponse experiments. Qual Eng 12(4):523–533
8. Gauri S, Pal S (2010) Comparison of performances of five prospective approaches for the multi-response optimization. Int J Adv Manuf Technol 48(12):1205–1220
9. Goh T (2009) Statistical thinking and experimental design as dual drivers of DFSS. Int J Six Sigma Compet Adv 5(1):2–9
10. Kazemzadeh R, Bashiri M, Atkinson A, Noorossana R (2008) A general framework for multiresponse optimization problems based on goal programming. Eur J Oper Res 189(2):421–429
11. Kim K, Lin D (2000) Simultaneous optimization of multiple responses by maximining exponential desirability functions. Appl Stat Ser C 49(3):311–325
12. Ko Y, Kim K, Jun C (2005) A new loss function-based method for multiresponse optimization. J Qual Tech 37(1):50–59
13. Lee M, Kim Y (2007) Separate response surface modeling for multiple response optimization: multivariate loss function approach. Int J Ind Eng 14(2):227–235
14. Mattson C, Messac A (2003) Concept selection using s-Pareto frontiers. AIAA J 41(6):1190–1198
15. Messac A, Sundararaj G, Tappeta R, Renaud J (2000) Ability of objective functions to generate points on non-convex pareto frontiers. AIAA J 38(6):1084–1091
16. Murphy T, Tsui K, Allen J (2005) A review of robust design methods for multiple responses. Res Eng Des 15(4):201–215
17. Myers R, Montgomery D (2002) Response surface methodology: process and product optimization using designed experiments, 2nd edn. Wiley, New Jersey
18. Myers R, Montgomery D, Anderson-Cook C (2009) Response surface methodology: process and product optimization using designed experiments, 3rd edn. Wiley, New York
19. Pignatiello J (1993) Strategies for robust multiresponse. IIE Trans 25(1):5–15
20. Shah H, Montgomery D, Matthew W (2004) Response surface modeling and optimization in multiresponse experiments using seemingly unrelated regressions. Qual Eng 16(3):387–397
21. Vining G (1998) A compromise approach to multiresponse optimization. J Qual Tech 30(4):309–313
22. Wu F, Chyu C (2004) Optimization of robust design for multiple quality characteristics. Int J Prod Res 42(2):337–354
23. Xu K, Lin D, Tang L, Xie M (2004) Multiresponse systems optimization using a goal attainment approach. IIE Trans 36(5):433–445
24. Yildiz A (2009) A new design optimization framework based on immune algorithm and Taguchi's method. Comput Ind 60(8):613–620
25. Younis A, Dong Z (2010) Trends, features, and tests of common and recently introduced global optimization methods. Eng Optim 42(8):691–718

Chapter 37
Inspection Policies in Service of Fatigued Aircraft Structures

Nicholas A. Nechval, Konstantin N. Nechval and Maris Purgailis

Abstract Fatigue is one of the most important problems of aircraft arising from their nature as multiple-component structures, subjected to random dynamic loads. For guaranteeing safety, the structural life ceiling limits of the fleet aircraft are defined from three distinct approaches: Safe-Life, Fail-Safe, and Damage Tolerance approaches. The common objectives to define fleet aircraft lives by the three approaches are to ensure safety while at the same time reducing total ownership costs. In this paper, the damage tolerance approach is considered and the focus is on the inspection scheme with decreasing intervals between inspections. The paper proposes an analysis methodology to determine appropriate decreasing intervals between inspections of fatigue-sensitive aircraft structures (as alternative to constant intervals between inspections often used in practice), so that risk of catastrophic accident during flight is minimized. The suggested approach is unique and novel in that it allows one to utilize judiciously the results of earlier inspections of fatigued aircraft structures for the purpose of determining the time of the next inspection and estimating the values of several parameters involved in the problem that can be treated as uncertain. Using in-service damage data and taking into account safety risk and maintenance cost at the same time, the above approach has been proposed to assess

N. A. Nechval (✉)
Department of Statistics, EVF Research Institute, University of Latvia,
Raina Blvd 19, Riga, LV-1050, Latvia
e-mail: nechval@junik.lv

K. N. Nechval
Department of Applied Mathematics, Transport and Telecommunication Institute,
Lomonosov Street 1, Riga, LV-1019, Latvia
e-mail: konstan@tsi.lv

M. Purgailis
Department of Cybernetics, University of Latvia, Raina Blvd 19, Riga, LV-1050, Latvia
e-mail: marispur@lanet.lv

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*,
Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_37,
© Springer Science+Business Media B.V. 2011

the reliability of aircraft structures subject to fatigue damage. An illustrative example is given.

37.1 Introduction

In spite of decades of investigation, fatigue response of materials is yet to be fully understood. This is partially due to the complexity of loading at which two or more loading axes fluctuate with time. Examples of structures experiencing such complex loadings are automobile, aircraft, off-shores, railways and nuclear plants. While most industrial failures involve fatigue, the assessment of the fatigue reliability of industrial components being subjected to various dynamic loading situations is one of the most difficult engineering problems. This is because material degradation processes due to fatigue depend upon material characteristics, component geometry, loading history and environmental conditions. The traditional analytical method of engineering fracture mechanics (EFM) usually assumes that crack size, stress level, material property and crack growth rate, etc. are all deterministic values which will lead to conservative or very conservative outcomes. However, according to many experimental results and field data, even in well-controlled laboratory conditions, crack growth results usually show a considerable statistical variability [1].

The analysis of fatigue crack growth is one of the most important tasks in the design and life prediction of aircraft fatigue-sensitive structures (for instance, wing, fuselage) and their components (for instance, aileron or balancing flap as part of the wing panel, stringer, etc.).

Several probabilistic or stochastic models have been employed to fit the data from various fatigue crack growth experiments. Among them, the Markov chain model [2], the second-order approximation model [3], and the modified second-order polynomial model [4]. Each of the models may be the most appropriate one to depict a particular set of fatigue growth data but not necessarily the others. All models can be improved to depict very accurately the growth data but, of course, it has to be at the cost of increasing computational complexity. Yang's model [3] and the polynomial model [4] are considered more appropriate than the Markov chain model [2] by some researchers through the introduction of a differential equation which indicates that fatigue crack growth rate is a function of crack size and other parameters. The parameters, however, can only be determined through the observation and measurement of many crack growth samples. Unfortunately, the above models are mathematically too complicated for fatigue researchers as well as design engineers. A large gap still needs to be bridged between the fatigue experimentalists and researchers who use probabilistic methods to study the fatigue crack growth problems.

Airworthiness regulations require proof that aircraft can be operated safely. This implies that critical components must be replaced or repaired before safety is

compromised. For guaranteeing safety, the structural life ceiling limits of the fleet aircraft are defined from three distinct approaches: Safe-Life, Fail-Safe, and Damage-Tolerant approaches. The common objectives to define fleet aircraft lives by the three approaches are to ensure safety while at the same time reducing total ownership costs. Although the objectives of the three approaches are the same, they vary with regard to the fundamental definition of service life.

The Safe-Life approach is based on the concept that significant damage, i.e. fatigue cracking, will not develop during the service life of a component. When the service life equals the design Safe-Life the component must be replaced. The Fail-Safe approach assumes initial damage as manufactured and its subsequent growth during service to detectable crack sizes or greater. Service life in Fail-Safe structures can thus be defined as the time to a service detectable damage. However, there are two major drawbacks to the Safe-Life and Fail-Safe approaches: (1) components are taken out of service even though they may have substantial remaining lives; (2) despite all precautions, cracks sometimes occur prematurely. These facts led the Airlines to introduce the Damage Tolerance approach, which is based on the concept that damage can occur and develop during the service life of a component. In this paper, the Damage Tolerance approach is considered and the focus is on the inspection scheme with decreasing intervals between inspections.

From an engineering standpoint the fatigue life of a component or structure consists of two periods: (i) crack initiation period, which starts with the first load cycle and ends when a technically detectable crack is present, and (ii) crack propagation period, which starts with a technically detectable crack and ends when the remaining cross section can no longer withstand the loads applied and fails statically. Periodic inspections of aircraft are common practice in order to maintain their reliability above a desired minimum level. The appropriate inspection intervals are determined so that the fatigue reliability of the entire aircraft structure remains above the minimum reliability level throughout its service life.

37.2 Inspection Scheme Under Fatigue Crack Initiation

At first, we consider in this section the problem of estimating the minimum time to crack initiation (warranty period or time to the first inspection) for a number of aircraft structure components, before which no cracks (that may be detected) in materials occur, based on the results of previous warranty period tests on the structure components in question. If in a fleet of k aircraft there are km of the same individual structure components, operating independently, the length of time until the first crack initially formed in any of these components is of basic interest, and provides a measure of assurance concerning the operation of the components in question. This leads to the consideration of the following problem. Suppose we have observations $X_1, \ldots, X_n$ as the results of tests conducted on the components; suppose also that there are km components of the same kind to be put into future

use, with times to crack initiation $Y_1, \ldots, Y_{\mathrm{km}}$. Then we want to be able to estimate, on the basis of $X_1, \ldots, X_n$, the shortest time to crack initiation $Y_{(1,\ \mathrm{km})}$ among the times to crack initiation $Y_1, \ldots, Y_{\mathrm{km}}$. In other words, it is desirable to construct lower simultaneous prediction limit, L_γ, which is exceeded with probability γ by observations or functions of observations of all k future samples, each consisting of m units. In this section, the problem of estimating $Y_{(1,\mathrm{km})}$, the smallest of all k future samples of m observations from the underlying distribution, based on an observed sample of n observations from the same distribution, is considered.

Assigning the time interval until the first inspection. Experiments show that the number of flight cycles (hours) at which a technically detectable crack will appear in a fatigue-sensitive component of aircraft structure follows the two-parameter Weibull distribution. The probability density function for the random variable X of the two-parameter Weibull distribution is given by

$$f(x|\beta,\delta) = \frac{\delta}{\beta}\left(\frac{x}{\beta}\right)^{\delta-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\delta}\right](x>0), \tag{37.1}$$

where $\delta > 0$ and $\beta > 0$ are the shape and scale parameters, respectively. The following theorem is used to assign the time interval until the first inspection (warranty period).

Theorem 1 (Lower one-sided prediction limit for the lth order statistic of the Weibull distribution). Let $X_1 < \cdots < X_r$ *be the first r ordered past observations from a sample of size n from the distribution* (37.1). *Then a lower one-sided conditional* $(1-\alpha)$ *prediction limit h on the lth order statistic* Y_l *of a set of m future ordered observations* $Y_1 < \cdots < Y_m$ *is given by*

$$\begin{aligned}
&\Pr\{Y_l > h|\mathbf{z}\} = \Pr\left\{\widehat{\delta}\ln\left(Y_l/\widehat{\beta}\right) > \widehat{\delta}\ln\left(h/\widehat{\beta}\right)|\mathbf{z}\right\} = \Pr\{W_l > w_h|\mathbf{z}\} \\
&\quad = \frac{\sum_{j=0}^{l-1}\left[\binom{l}{j}(-1)^{l-1-j}\frac{1}{m-j}\int_0^\infty v^{r-2}e^{v\widehat{\delta}\sum_{i=1}^{r}\ln\left(x_i/\widehat{\beta}\right)}\left((m-j)e^{vw_h} + \sum_{i=1}^{r}e^{v\widehat{\delta}\ln\left(x_i/\widehat{\beta}\right)} + (n-r)e^{v\widehat{\delta}\ln\left(x_r/\widehat{\beta}\right)}\right)^{-r}dv\right]}{\sum_{j=0}^{l-1}\left[\binom{l}{j}(-1)^{l-1-j}\frac{1}{m-j}\int_0^\infty v^{r-2}e^{v\widehat{\delta}\sum_{i=1}^{r}\ln\left(x_i/\widehat{\beta}\right)}\left(\sum_{i=1}^{r}e^{v\widehat{\delta}\ln\left(x_i/\widehat{\beta}\right)} + (n-r)e^{v\widehat{\delta}\ln\left(x_r/\widehat{\delta}\right)}\right)^{-r}dv\right]} \\
&\quad = 1-\alpha,
\end{aligned} \tag{37.2}$$

where $\widehat{\beta}$ *and* $\widehat{\delta}$ *are the maximum likelihood estimators of* β *and* δ *based on the first r ordered past observations* $(X_1,\ldots, X_r)$ *from a sample of size n from the Weibull distribution, which can be found from solution of*

$$\widehat{\beta} = \left(\left[\sum_{i=1}^{r} x_i^{\widehat{\delta}} + (n-r) x_r^{\widehat{\delta}} \right] \Big/ r \right)^{1/\widehat{\delta}}, \tag{37.3}$$

and

$$\widehat{\delta} = \left[\left(\sum_{i=1}^{r} x_i^{\widehat{\delta}} \ln x_i + (n-r) x_r^{\widehat{\delta}} \ln x_r \right) \left(\sum_{i=1}^{r} x_i^{\widehat{\delta}} + (n-r) x_r^{\widehat{\delta}} \right)^{-1} - \frac{1}{r} \sum_{i=1}^{r} \ln x_i \right]^{-1}, \tag{37.4}$$

$$\mathbf{z} = (z_1, z_2, \ldots, z_{r-2}), \quad Z_i = \widehat{\delta} \ln\left(X_i / \widehat{\beta} \right), \quad i = 1, \ldots, r-2, \tag{37.5}$$

$$W_l = \widehat{\delta} \ln(Y_l / \widehat{\beta}), \quad w_h = \widehat{\delta} \ln(h / \widehat{\beta}). \tag{37.6}$$

(*Observe that an upper one-sided conditional* α *prediction limit h on the lth order statistic* Y_l *may be obtained from a lower one-sided conditional* $(1-\alpha)$ *prediction limit by replacing* $1 - \alpha$ *by* α.)

Proof The proof is given by Nechval et al. [5] and so it is omitted here. □

Corollary 1.1 *A lower one-sided conditional* $(1 - \alpha)$ *prediction limit h on the minimum* Y_1 *of a set of m future ordered observations* $Y_1 \leq \cdots \leq Y_m$ *is given by*

$$\Pr\{Y_1 > h|\mathbf{z}\} = \Pr\left\{ \widehat{\delta} \ln\left(Y_1 \Big/ \widehat{\beta} \right) > \widehat{\delta} \ln\left(h \Big/ \widehat{\beta} \right) | \mathbf{z} \right\} = \Pr\{W_1 > w_h|\mathbf{z}\}$$

$$= \frac{\int_0^\infty v^{r-2} e^{v\widehat{\delta} \sum_{i=1}^{r} \ln\left(x_i/\widehat{\beta}\right)} \left(\mathrm{me}^{v w_h} + \sum_{i=1}^{r} e^{v\widehat{\delta} \ln\left(x_i/\widehat{\beta}\right)} + (n-r) e^{v\widehat{\delta} \ln\left(x_r/\widehat{\beta}\right)} \right)^{-r} dv}{\int_0^\infty v^{r-2} e^{v\widehat{\delta} \sum_{i=1}^{r} \ln\left(x_i/\widehat{\beta}\right)} \left(\sum_{i=1}^{r} e^{v\widehat{\delta} \ln\left(x_i/\widehat{\beta}\right)} + (n-r) e^{v\widehat{\delta} \ln\left(x_r/\widehat{\beta}\right)} \right)^{-r} dv}$$

$$= 1 - \alpha. \tag{37.7}$$

Thus, when $l = 1$ (37.2) *reduces to formula* (37.7).

Theorem 2 (Lower one-sided prediction limit for the *l*th order statistic of the exponential distribution). *Under conditions of Theorem* 1, *if* $\delta = 1$, *we deal with the exponential distribution, the probability density function of which is given by*

$$f(x|\beta) = \frac{1}{\beta} \exp\left(-\frac{x}{\beta} \right) (x > 0). \tag{37.8}$$

Then a lower one-sided conditional $(1-\alpha)$ *prediction limit* h *on the* lth *order statistic* Y_l *of a set of* m *future ordered observations* $Y_1 < \cdots < Y_m$ *is given by*

$$\Pr\{Y_l \geq h|S_\beta = s_\beta\} = \Pr\left\{\frac{Y_l}{S_\beta} \geq \frac{h}{s_\beta}|S_\beta = s_\beta\right\}$$
$$= \Pr\{W_l > w_h\} = \frac{1}{\mathrm{B}(l, m-l+1)} \sum_{j=0}^{l-1} \binom{l-1}{j} (-1)^j$$
$$\times \frac{1}{(m-l+1+j)[1+w_l(m-l+1+j)]^r} = 1-\alpha. \tag{37.9}$$

where

$$W_l = \frac{Y_l}{S_\beta}, \quad w_h = \frac{h}{s_\beta}, \quad S_\beta = \sum_{i=1}^{r} X_i + (m-r)X_r. \tag{37.10}$$

Proof It follows readily from standard theory of order statistics that the distribution of the lth order statistic Y_l from a set of m future ordered observations $Y_1 \leq \cdots \leq Y_m$ is given by

$$f(y_l|\beta)dx_l = \frac{1}{\mathrm{B}(l, m-l+1)}[F(x_l|\beta)]^{l-1}[1-F(x_l|\beta)]^{m-l}dF(x_l|\beta), \tag{37.11}$$

where

$$F(x|\beta) = 1 - \exp(-x/\beta). \tag{37.12}$$

The factorization theorem gives

$$S_\beta = \sum_{i=1}^{r} X_i + (n-r)X_r \tag{37.13}$$

sufficient for β. The density of S_β is given by

$$g(s_\beta|\beta) = \frac{1}{\Gamma(r)\beta^r} s_\beta^{r-1} \exp\left(-\frac{s_\beta}{\beta}\right), \quad s_\beta \geq 0. \tag{37.14}$$

Since Y_l, S_β are independent, we have the joint density of Y_l and S_β as

$$f(y_l, s_\beta|\beta) = \frac{1}{\mathrm{B}(l, m-l+1)} \frac{1}{\Gamma(r)} [1-e^{-x_l/\beta}]^{l-1} [e^{-x_l/\beta}]^{m-l+1} \frac{1}{\beta^{r+1}} s_\beta^{r-1} e^{-s_\beta/\beta}. \tag{37.15}$$

Making the transformation $w_l = y_l/s_\beta$, $s_\beta = s_\beta$, and integrating out s_β, we find the density of W_l as the beta density

$$f(w_l) = \frac{r}{\mathrm{B}(l, m-l+1)} \sum_{j=0}^{l-1} \binom{l-1}{j} (-1)^j$$
$$\times \frac{1}{[(m-l+1+j)w_l + 1]^{r+1}}, \ 0 < w_l < \infty. \tag{37.16}$$

This ends the proof. □

Corollary 2.1 *A lower one-sided conditional* $(1-\alpha)$ *prediction limit* h *on the minimum* Y_1 *of a set of* m *future ordered observations* $Y_1 \leq \cdots \leq Y_m$ *is given by*

$$\begin{aligned}\Pr\{Y_1 \geq h | S_\beta = s_\beta\} &= \Pr\{Y_1/S_\beta \geq h/s_\beta | S_\beta = s_\beta\} \\ &= \Pr\{W_1 > w_h\} = 1/(1+mw_h)^r = 1-\alpha. \end{aligned} \tag{37.17}$$

Example Consider the data of fatigue tests on a particular type of structural components (stringer) of aircraft IL-86. The data are for a complete sample of size $r = n = 5$, with observations of time to crack initiation (in number of 10^4 flight hours): $X_1 = 5$, $X_2 = 6.25$, $X_3 = 7.5$, $X_4 = 7.9$, $X_5 = 8.1$.

Goodness-of-fit testing. It is assumed that X_i, $i = 1(1)5$, follow the two-parameter Weibull distribution (37.1), where the parameters β and δ are unknown. We assess the statistical significance of departures from the Weibull model by performing empirical distribution function goodness-of-fit test. We use the S statistic (Kapur and Lamberson [6]). For censoring (or complete) datasets, the S statistic is given by

$$S = \frac{\sum_{i=[r/2]+1}^{r-1}\left(\frac{\ln(x_{i+1}/x_i)}{M_i}\right)}{\sum_{i=1}^{r-1}\left(\frac{\ln(x_{i+1}/x_i)}{M_i}\right)} = \frac{\sum_{i=3}^{4}\left(\frac{\ln(x_{i+1}/x_i)}{M_i}\right)}{\sum_{i=1}^{4}\left(\frac{\ln(x_{i+1}/x_i)}{M_i}\right)} = 0.184, \tag{37.18}$$

where $[r/2]$ is a largest integer $\leq r/2$, the values of M_i are given in Table 13 (Kapur and Lamberson [6]). The rejection region for the α level of significance is $\{S > S_{n;\alpha}\}$. The percentage points for $S_{n;\alpha}$ were given by Kapur and Lamberson [6]. For this example,

$$S = 0.184 < S_{n=5;\alpha=0.05} = 0.86. \tag{37.19}$$

Thus, there is not evidence to rule out the Weibull model. The maximum likelihood estimates of the unknown parameters β and δ are $\widehat{\beta} = 7.42603$ and $\widehat{\delta} = 7.9081$, respectively.

Warranty period estimation. It follows from (37.7) that

$$\begin{aligned}\Pr\{Y_1 > h|\mathbf{z}\} &= \Pr\left\{\widehat{\delta}\ln\left(Y_1/\widehat{\beta}\right) > \widehat{\delta}\ln\left(h/\widehat{\beta}\right)|\mathbf{z}\right\} = \Pr\{W_1 > w_h|\mathbf{z}\} \\ &= \Pr\{W_1 > -8.4378;\ \mathbf{z}\} = 0.0000141389/0.0000148830 = 0.95 \end{aligned} \tag{37.20}$$

and a lower 0.95 prediction limit for Y_1 is $h = 2.5549$ ($\times 10^4$) flight hours, i.e., we have obtained the time interval until the first inspection (or warranty period) equal to 25,549 flight hours with confidence level $\gamma = 1 - \alpha = 0.95$.

Inspection Policy after Warranty Period. Let us assume that in a fleet of m aircraft there are m of the same individual structure components, operating independently. Suppose an inspection is carried out at time τ_j, and this shows that initial crack (which may be detected) has not yet occurred. We now have to

schedule the next inspection. Let Y_1 be the minimum time to crack initiation in the above components. In other words, let Y_1 be the smallest observation from an independent second sample of m observations from the distribution (37.1). Then the inspection times can be calculated (from (37.23) using (37.22)) as

$$\tau_j = \widehat{\beta}\exp(w_{\tau_j}/\widehat{\delta}), \quad j \geq 1, \tag{37.21}$$

where it is assumed that $\tau_0 = 0$, τ_1 is the time until the first inspection (or warranty period), $w_{\tau_j}$is determined from

$$\begin{aligned}
&\Pr\{Y_1 > \tau_j | Y_1 > \tau_{j-1}, \mathbf{z}\} \\
&= \Pr\left\{\widehat{\delta}\ln\left(Y_1/\widehat{\beta}\right) > \widehat{\delta}\ln\left(\tau_j/\widehat{\beta}\right) \middle| \widehat{\delta}\ln\left(Y_1/\widehat{\beta}\right) > \widehat{\delta}\ln\left(\tau_{j-1}/\widehat{\beta}\right), \mathbf{z}\right\} \\
&= \Pr\{W_1 > w_{\tau_j} | W_1 > w_{\tau_{j-1}}, \mathbf{z}\} = \Pr\{\mathbf{W_1} > \mathbf{w_{\tau_j}} | \mathbf{z}\} / \Pr\{\mathbf{W_1} > \mathbf{w_{\tau_{j-1}}} | \mathbf{z}\} = \mathbf{1} - \alpha,
\end{aligned} \tag{37.22}$$

where

$$W_1 = \widehat{\delta}\ln\left(Y_1/\widehat{\beta}\right), \quad w_{\tau_j} = \widehat{\delta}\ln\left(\tau_j/\widehat{\beta}\right), \tag{37.23}$$

$\widehat{\beta}$ and $\widehat{\delta}$ are the MLE's of β and δ, respectively, and can be found from solution of (37.3) and (37.4), respectively.

It will be noted that if $\delta = 1$, then it follows from (37.10) that

$$\tau_j = w_{\tau_j} s_\beta, \quad j \geq 1, \tag{37.24}$$

where w_{τ_j} is determined from

$$\begin{aligned}
&\Pr\{Y_1 > \tau_j | Y_1 > \tau_{j-1}, S_\beta = s_\beta\} = \Pr\left\{\frac{Y_1}{S_\beta} > \frac{\tau_j}{s_\beta} \middle| \frac{Y_1}{S_\beta} > \frac{\tau_{j-1}}{s_\beta}, S_\beta = s_\beta\right\} \\
&= \Pr\{W_1 > w_{\tau_j} | W_1 > w_{\tau_{j-1}}\} = \frac{\Pr\{W_1 > w_{\tau_j}\}}{\Pr\{W_1 > w_{\tau_{j-1}}\}} = \frac{1/(1+mw_{\tau_j})^r}{1/(1+mw_{\tau_{j-1}})^r} = 1 - \alpha,
\end{aligned} \tag{37.25}$$

$$W_1 = \frac{Y_1}{S_\beta}, \quad w_{\tau_j} = \frac{\tau_j}{s_\beta}, \quad S_\beta = \sum_{i=1}^{r} X_i + (m-r)X_r. \tag{37.26}$$

If it is assumed that $\tau_0 = 0$, then the time until the first inspection (or warranty period) τ_1 is given by

$$\tau_1 = w_{\tau_1} s_\beta, \tag{37.27}$$

where $w_{\tau_1}$is determined from (37.25) as

$$w_{\tau_1} = \arg[1/(1+mw_{\tau_1})^r = 1 - \alpha]. \tag{37.28}$$

Table 37.1 Inspection time sequence

$w_{\tau_j} \equiv w_j$	Inspection time τ_j ($\times 10^4$ flight hours)	Interval $\tau_{j+1} - \tau_j$ (flight hours)
–	$\tau_0 = 0$	–
$w_1 = -8.4378$	$\tau_1 = 2.5549$	25,549
$w_2 = -6.5181$	$\tau_2 = 3.2569$	7,020
$w_3 = -5.5145$	$\tau_3 = 3.6975$	4,406
$w_4 = -4.8509$	$\tau_4 = 4.0212$	3,237
$w_5 = -4.3623$	$\tau_5 = 4.2775$	2,563
$w_6 = -3.9793$	$\tau_6 = 4.4898$	2,123
$w_7 = -3.6666$	$\tau_7 = 4.6708$	1,810
$w_8 = -3.4038$	$\tau_8 = 4.8287$	1,579
$w_9 = -3.1780$	$\tau_9 = 4.9685$	1,398
⋮	⋮	⋮

It follows from (37.25) also that

$$\Pr\{W_1 > w_{\tau_j}\} = 1/(1 + mw_{\tau_j})^r = (1 - \alpha)^j. \tag{37.29}$$

Thus, we have from (37.24) and (37.29) that

$$\tau_j = \frac{s_\beta}{m}[(1 - \alpha)^{-j/r} - 1], \quad j \geq 1. \tag{37.30}$$

But again, for instance, consider the data of fatigue tests on a particular type of structural components of aircraft IL-86: $x_1 = 5$, $x_2 = 6.25$, $x_3 = 7.5$, $x_4 = 7.9$, $x_5 = 8.1$ (in number of 10^4 flight hours) given above, where $r = n = 5$ and the maximum likelihood estimates of unknown parameters β and δ are $\widehat{\beta} = 7.42603$ and $\widehat{\delta} = 7.9081$, respectively. Thus, using (37.21) and (37.22) with $\tau_1 = 2.5549$ ($\times 10^4$ flight hours) (the time of the first inspection), we have obtained the following inspection time sequence (see Table 37.1).

Optimization of inspection policy. Consider the case where an optimal inspection policy has to be computed with linear costs. Let c_1 be the cost of each of the inspections. If crack occurs at time t and is detected at the jth inspection time τ_j, so that $\tau_j \geq t$, let the cost due to undetected crack be $c_2(\tau_j - t)$. It will be noted that under small failure probability one can be restricted by the first term of the Taylor series as a presentation of a proper loss function $c_2(.)$. Then an optimal inspection policy is one, which minimizes the expected value of the total cost

$$C = c_1 j + c_2(\tau_j - t). \tag{37.31}$$

Taking into account (37.13), we obtain for the Weibull case:

$$
\begin{aligned}
E\{C|z\} &= c_1 \sum_{k=0}^{\infty} k \Pr\{j=k\} + c_2 \left(\sum_{j=1}^{\infty} \tau_j \Pr\{\tau = \tau_j\} - E\{T|z\} \right) \\
&= c_1 \sum_{k=1}^{\infty} k[\Pr\{j > k-1\} - \Pr\{j > k\}] + c_2 \left(\sum_{j=1}^{\infty} \tau_j [1-\alpha]^{j-1} \alpha - E\{T|z\} \right) \\
&= c_1 \sum_{k=0}^{\infty} \Pr\{j > k\}] + c_2 \left(\sum_{j=1}^{\infty} \tau_j [1-\alpha]^{j-1} \alpha - E\{T|z\} \right) \\
&= c_1 \sum_{k=0}^{\infty} (1-\alpha)^k + c_2 \left(\sum_{j=1}^{\infty} \tau_j [1-\alpha]^{j-1} \alpha - E\{T|z\} \right) \\
&= \frac{c_1}{\alpha} + c_2 \left(\sum_{j=1}^{\infty} \widehat{\sigma} \exp\left(\frac{w_{\tau_j}}{\widehat{\delta}} \right) [1-\alpha]^{j-1} \alpha - E\{T|z\} \right),
\end{aligned} \tag{37.32}
$$

Now an optimal α has to be found such that minimizes (37.32). The optimal value of α has to be determined numerically. Then, the optimal $\tau_j, j \geq 1$, are found from (37.23) using (37.22).

37.3 Inspection Scheme Under Fatigue Crack Propagation

Probabilistic model of fatigue crack growth Many probabilistic models of fatigue crack growth are based on the deterministic crack growth equations. The most well known equation is

$$
\frac{\mathrm{d}a(t)}{\mathrm{d}t} = q(a(t))^b \tag{37.33}
$$

in which q and b are constants to be evaluated from the crack growth observations. The independent variable t can be interpreted as stress cycles, flight hours, or flights depending on the applications [7]. It is noted that the power-law form of $q(a(t))^b$ at the right hand side of (37.33) can be used to fit some fatigue crack growth data appropriately and is also compatible with the concept of Paris–Erdogan law. The service time for a crack to grow from size $a(t_0)$ to $a(t)$ (where $t > t_0$) can be found by performing the necessary integration

$$
\int_{t_0}^{t} dt = \int_{a(t_0)}^{a(t)} \frac{\mathrm{d}v}{qv^b} \tag{37.34}
$$

to obtain

$$
t - t_0 = \frac{[a(t_0)]^{-(b-1)} - [a(t)]^{-(b-1)}}{q(b-1)}. \tag{37.35}
$$

In this paper, we consider a stochastic version of (37.35),

$$\frac{1}{a_0^{b-1}} - \frac{1}{a^{b-1}} = (b-1)q(t-t_0) + V, \tag{37.36}$$

where $a_0 \equiv a(t_0)$, $a \equiv a(t)$ Let us assume that $V \sim \mathrm{N}(0,[(b\text{-}1)\sigma(t-t_0)^{1/2}]^2)$, then the probability that crack size $a(t)$ will exceed any given (say, maximum allowable) crack size $a^\bullet$ can be derived and expressed as

$$\Pr\{a(t) \geq a^\bullet\} = 1 - \Phi\left(\left[\frac{(a_0^{-(b-1)} - (a^\bullet)^{-(b-1)}) - (b-1)q(t-t_0)}{(b-1)\sigma(t-t_0)^{1/2}}\right]\right), \tag{37.37}$$

where $\Phi(.)$ is the standard normal distribution function. In this case, the conditional probability density function of a is given by

$$f(a,t|b,q,\sigma) = \frac{a^{-b}}{\sigma[2\pi(t-t_0)]^{1/2}} \exp\left(-\frac{1}{2}\left[\frac{(a_0^{-(b-1)} - a^{-(b-1)}) - (b-1)q(t-t_0)}{(b-1)\sigma(t-t_0)^{1/2}}\right]^2\right) \tag{37.38}$$

This model allows one to characterize the random properties that vary during crack growth [8–10].

Inspection policy under parametric certainty of the model Let us assume that all the parameters of the crack exceedance probability (37.37) are known. Then the inspection times can be calculated recursively from

$$\Pr\{a(\tau_j) < a^\bullet | a(\tau_{j-1}) < a^\bullet\} = \frac{\Pr\{a(\tau_j) < a^\bullet\}}{\Pr\{a(\tau_{j-1}) < a^\bullet\}} = 1 - \alpha, \quad j \geq 1, \tag{37.39}$$

where

$$\Pr\{a(\tau_j) < a^\bullet\} = \Phi\left(\left[\frac{(a_0^{-(b-1)} - (a^\bullet)^{-(b-1)}) - (b-1)q(\tau_j - \tau_0)}{(b-1)\sigma(\tau_j - \tau_0)^{1/2}}\right]\right), \tag{37.40}$$

$\tau_0 = 0$, τ_1 is the time of the inspection when the initial crack was detected. It is assumed that cracks start growing from the time the aircraft entered service. For typical aircraft metallic materials, an initial discontinuity size (a_0) found through quantitative fractography is approximately between 0.02 and 0.05 mm. Choosing a typical value for initial discontinuity state (e.g., 0.02 mm) is more conservative than choosing an extreme value (e.g., 0.05 mm). This implies that if the lead cracks can be attributed to unusually large initiating discontinuities then the available life increases.

Inspection policy under parametric uncertainty of the model Let us assume that the parameters b, q and σ of the crack exceedance probability (37.37) are unknown. Given the data describing a single crack, say a sequence $\{(a_j, \tau_j)\}_{j=1}^n$, it is easy to construct a log-likelihood using the density given by (37.28) and estimate the parameters b, q and σ by maximum likelihood. The log-likelihood is

$$L(b,q,\sigma|\{(a_j,\tau_j)\} = -b\sum_{j=1}^{n}\ln a_j - n\ln\sigma$$
$$-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{a_0^{1-b}-a_j^{1-b}-(b-1)q(\tau_j-\tau_0)}{(b-1)\sigma(\tau_j-\tau_0)^{1/2}}\right)^2. \tag{37.41}$$

Inspection shows that this differs from the standard least-squares equation only in the term $-b\sum\ln a$, where the subscript i has been dropped. The likelihood estimators are obtained by solving the equations

$$dL/db = 0; \quad dL/dq = 0; \quad dL/d\sigma = 0. \tag{37.42}$$

In this case the equations have no closed solution. However, it is easy to see that the estimators for q and σ given b are the usual least-squares estimators for the coefficients in (37.42) conditioned on b,

$$\hat{q}(b) = \frac{1}{b-1}\left(na_0^{1-b} - \sum_{j=1}^{n}a_j^{1-b}\right)\left(\sum_{j=1}^{n}(\tau_j-\tau_0)\right)^{-1}, \tag{37.43}$$

$$\hat{\sigma}^2(b) = \frac{1}{n(b-1)^2}\sum_{j=1}^{n}\frac{[a_0^{1-b}-a_j^{1-b}-\hat{q}(b)(b-1)(\tau_j-\tau_0)]^2}{\tau_j-\tau_0}, \tag{37.44}$$

and on substituting these back in the log-likelihood gives a function of b alone,

$$L(b) = -b\sum_{j=1}^{n}\ln a_j - n\ln[\hat{\sigma}(b)] - n/2. \tag{37.45}$$

Thus the technique is to search for the value of b that maximizes $L(b)$ by estimating q and σ as functions of b and substituting in $L(b)$. In this study a simple golden-section search worked very effectively. It will be noted that if we deal with small sample of the data describing a single crack, say a sequence $\{(a_j,\tau_j)\}_{j=1}^{n}$, then the estimates of the unknown parameters b, q and σ can be obtained via the Generalized Likelihood Ratio Test as follows. Let us assume (without loss of generality) that there are available only two past samples of the data describing a single similar crack, say sequences $\{(a_j^{(1)},\tau_j^{(1)})\}_{j=1}^{n_1}$ and $\{(a_j^{(2)},\tau_j^{(2)})\}_{j=1}^{n_2}$ with the unknown parameters (b_1, q_1, σ_1) and (b_2, q_2, σ_2), respectively, where n_1, $n_2 > n$. Then the likelihood ratio statistic for testing the null hypothesis H_1: ($b = b_1$, $q = q_1$, $\sigma = \sigma_1$) versus the alternative hypothesis H_2: ($b = b_2$, $q = q_2$, $\sigma = \sigma_2$) is given by

$$\mathrm{LR} = \frac{\max_{H_1}\prod_{j=1}^{n}f(a_j,\tau_j|b_1,q_1,\sigma_1)\prod_{i=1}^{2}\prod_{j=1}^{n_i}f(a_j^{(i)},\tau_j^{(i)}|b_i,q_i,\sigma_i)}{\max_{H_2}\prod_{j=1}^{n}f(a_j,\tau_j|b_2,q_2,\sigma_2)\prod_{i=1}^{2}\prod_{j=1}^{n_i}f(a_j^{(i)},\tau_j^{(i)}|b_i,q_i,\sigma_i)}, \tag{37.46}$$

and hypothesis H_1 or H_2 is favoured according to whether LR is greater or less than 1, i.e.

$$\mathrm{LR}\begin{cases} > 1, & \text{then } H_1\ (\widehat{b} = \widehat{b}_1,\ \widehat{q} = \widehat{q}_1,\ \widehat{\sigma} = \widehat{\sigma}_1) \\ \leq 1, & \text{then } H_2\ (\widehat{b} = \widehat{b}_2,\ \widehat{q} = \widehat{q}_2,\ \widehat{\sigma} = \widehat{\sigma}_2). \end{cases} \tag{37.47}$$

The parametric estimates obtained after each inspection are treated as if they were the true values in order to obtain from (37.39) an adaptive inspection time sequence.

37.4 Conclusion and Future Work

This paper suggests innovative frequentist (non-Bayesian) statistical approach to planning warranty period and in-service inspections for fatigue-sensitive aircraft structure components under parametric uncertainty of the underlying lifetime distributions and fatigue reliability requirement. This is very important since decision rules, which are optimal in the absence of uncertainty, need not even be approximately optimal in the presence of such uncertainty. The paper presents more accurate statistical procedures for solving the fatigue reliability problems, which are attractively simple and easy to apply in practice for situations where it is difficult to quantify the costs associated with inspections and undetected cracks.

The results obtained in this work can be used to solve the service problems of the following important engineering structures: (1) transportation systems and vehicles—aircraft, space vehicles, trains, ships; (2) civil structures—bridges, dams, tunnels; (3) power generation—nuclear, fossil fuel and hydroelectric plants; (4) high-value manufactured products—launch systems, satellites, semiconductor and electronic equipment; (5) industrial equipment – oil and gas exploration, production and processing equipment, chemical process facilities, pulp and paper.

Acknowledgments This research was supported in part by Grant No. 07.2036, Grant No. 09.1014, and Grant No. 09.1544 from the Latvian Council of Science and the National Institute of Mathematics and Informatics of Latvia.

References

1. Nechval KN, Nechval NA, Bausova I, Skiltere D, Strelchonok VF (2006) Prediction of fatigue crack growth process via artificial neural network technique. Int J Comput 5:21–32
2. Bogdanoff JL, Kozin F (1985) Probabilistic models of cumulative damage. Wiley, New York
3. Yang JN, Manning SD (1996) A simple second order approximation for stochastic crack growth analysis. Eng Fract Mech 53:677–686
4. Wu WF, Ni CC, Liou HY (2001) Random outcome and stochastic analysis of some fatigue crack growth data. Chin J Mech 17:61–68

5. Nechval NA, Nechval KN, Berzinsh G, Danovich V, Purgailis M (2008) Prediction limits for order statistics in future samples with some applications to product lifecycle. In: Horvath I, Rusak Z (eds) Tools and methods of competitive engineering, 2nd edn. vol 2. Delft University of Technology, The Netherlands, pp 881–894
6. Kapur KC, Lamberson LR (1977) Reliability in Engineering Design. Wiley, New York
7. Nechval KN, Nechval NA, Berzins G, Purgailis M, Rozevskis U, Strelchonok VF (2009) Optimal adaptive inspection planning process in service of fatigued aircraft structures. In: Al-Begain K, Fiems D, Horvath G (eds) LNCS, vol 5513. Springer, Berlin, pp 354–369
8. Nechval NA, Nechval KN, Berzins G, Purgailis M, Rozevskis U (2008) Stochastic fatigue models for efficient planning inspections in service of aircraft structures. In: Al-Begain K, Heindl A, Telek M (eds) LNCS, vol 5055. Springer, Berlin, pp 114–127
9. Nechval NA, Nechval KN (2008) Statistical identification of an observable process. Comput Modell New Technol 12:38–46
10. Nechval NA, Purgailis M, Cikste K, Nechval KN, Planning inspections of fatigued aircraft structures via damage tolerance approach. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, pp 2470–2475

Chapter 38
Toxicokinetic Analysis of Asymptomatic Hazard Profile of Welding Fumes and Gases

Joseph I. Achebo and Oviemuno Oghoore

Abstract In this research paper, 15 welders with about 1–20 years working experience in a welding firm, and who had also been diagnosed with fume related illnesses, were investigated. The fumes generated from E6010 electrodes using the shielded metal arc welding (SMAW) process, were collected and analyzed. A fume formation rate of 0.195 g/min was obtained under normal operating conditions and the average size of the agglomerated particles was found to be 2.14 μm. Such fine particles would easily settle within the welders' lungs; and the minute morphology of these fume particles, makes them deleterious to health. It was discovered that the critical time frame of 8.1–13.3 years, was the expected time range within which the welders were likely to begin to have symptoms. Crucially, it was determined that hazard rate is proportional to the expected time within which a welder actually becomes ill. Hence it is recommended that the evolution of toxic gases should be controlled at source. This study has comprehensively considered the hazard profile of welding fumes and gases which have evolved during the SMAW process.

38.1 Introduction

This research was based on an investigation of a welding and metal fabrication firm, specializing in the manufacture of overhead and underground petroleum storage tanks, with particular focus on the harmful effects of the welders' long

J. I. Achebo (✉) · O. Oghoore
Department of Production Engineering, University of Benin, Benin City, EdoState, Nigeria
e-mail: josephachebo@yahoo.co.uk

O. Oghoore
e-mail: oviemuno2002@yahoo.co.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_38,
© Springer Science+Business Media B.V. 2011

term exposure to welding fumes. The preferred welding method for joining the various pieces of metal is the shielded metal arc welding (SMAW) process, utilizing flux coated E6010 electrodes. The shielded metal arc welding process utilizes a consumable electrode coated in flux to make a multilayer of weld. This happens when an electric current is used to generate an electric arc. The arc's intense heat causes an explosion of the flux coating of the electrode, which disintegrates and forms vapor. The molten metal of the electrode, under intense arc heat becomes liquid, and forms metal droplets. As the arc temperature is sustained the droplets detach from the electrode tip and as the arc temperature is increasingly sustained, the detaching liquid metal globule explodes and evaporates into the atmosphere. Due to convection, where the cool air displaces hot air, the vaporized metal condenses into minuscule sized particle invisible to the naked eye, and the smallest particles or spatter persist as fumes. Welders inhale the floating condensed particulates and toxic gases and these particulates are very harmful to health, leading to serious lung injury, depending on the welders' level of exposure. This is the reason, some researchers say that welding poses a serious threat to health and safety [1]. Even though the fumes may sometimes be invisible, and would often be dispersed with the passage of time, through diffusion, traces of the deleterious vapor may still persist on site thereby posing a stealth-like danger to welders and non-welders alike, within the work environment.

For most firms, profit maximization, often takes preeminence over the health and safety of employees. Consequently, over the last few decades international organizations such as the International Labour Organization (ILO) and other watchdog labour bodies began to canvass for improved workplace health and safety. This has lead to the emergence of more stringent laws, such as compulsory medical insurance and industrial accidents compensation packages which have dramatically increased production costs for large and small industrial firms alike. Today, more and more industrial firms now focus on reducing industrial accidents and injury by studying and improving workplace health and safety. While many of the workplace injuries are apparent, requiring immediate medical attention and hospitalization, there are other workplace injuries that are latent and asymptomatic. Thus, if over the years a welder inhales these gases, vapors, and fumes, in quantities greater than those set by the occupational exposure limits (OELs), it is likely that welders' health will suffer. This observation classifies the health problems caused by welding fumes as Asymptomatic.

Achebo and Oghoore [2] were of the opinion that asymptomatic hazards are associated with unexpected or latent human errors. The error in this case is the consistent exposure of the welders and other personnel to deleterious welding fumes. Over the passage of years these injuries and the attendant insurance cover and workmen's compensation could overwhelm production costs.

Achebo and Oghoore [2] wrote that a hazard rate deals with the probability of exposing an employee to risk posed by the work environment and the level of danger such risks could have on him. This research thereafter focuses on the issue of the fume formation rate, which determines the amount of fume generated during any welding operation. Also, since it is already established that the elements of

these fume constitute a serious hazard to welders, the research developed a hazard profile based on the said effect of the fume constituents on welders, using probability failure functions.

The purpose of this paper is to investigate the variables relating to the effects of these fumes on the workers of the firm; to subject data to statistical and mathematical computation and analysis. In summary, this is a toxicokinetic analysis of asymptomatic hazard profile of welding fumes and gases.

38.2 Fume Composition Analysis

Chan et al. [3] said that fumes are generated from either the base metal, filler metal, the electrode coatings, or the shielding gas; but Stern [4] said that as much as 95% of the welding fumes in SMAW come from the electrodes. Hilton and Plumridge [5] and Pires et al. [6] observed that the fumes produced during gas metal arc welding (GMAW) are composed of oxides and metal vapors that originate predominately from the welding wire/electrode, while the base metal usually contributes <10% to the total mass of fumes. Nederman [1] wrote that around 90% of welding fumes originates from the consumable filler wire or flux coated electrode, with the base metal making only a small contribution. This claim is also supported by Lyttle [7] and Spear [8] who said that about 90–95% of the fumes are generated from the filler metal and flux coating/core of consumable electrodes. Chen et al. [3] were of the opinion that the chemical constituents of the welding fumes play an important role in determining the hazardous nature of the welding environment. Also important from the toxicity standpoint are the size and concentration of the particles within the weld fume.

The fumes contain all the elements present in the consumable electrode, but often in very different proportions. Volatile components have a higher concentration in the fumes than in the consumables. This is because the components possessing higher melting points do not readily ignite, and therefore tend to have a lower concentration in the fume.

Flux coated mild steel electrode E6010 wires, 3.0 mm in diameter, and 350 mm in length were used in this study; their chemical compositions are shown in Tables 38.1 and 38.2. The constituent elements of the weld deposits were determined in their various proportions (percent by weight). The chemical composition of the base metal and the corresponding all weld metal are found in Table 38.1.

The process parameters used to obtain the weld metal deposits are listed in Table 38.3. The chemical composition of the condensed fumes was analyzed using a spectrometer.

The fumes used for this investigation were collected from an exhaust fitter fixed to an enclosed welding chamber. The welders were properly protected from the welding fumes using certified respiratory equipment. The trapped fumes were collected and analyzed by using the X-ray fluorescence technique. This technique has been explained by Chan et al. [3]. The chemical composition of fumes is given

Table 38.1 Chemical composition of analysis

Element	%C	%Mn	%Si	%Mo	%Cr	%Ni	%Al	%Cu	%Ti	%P	%S	%O	%N	%Fe
Filler Metal	0.05	1.89	0.09	0.31	0.08	0.18	0.02	0.47	0.04	0.021	0.012	0.004	0.006	96.824
Base Metal	0.21	1.42	0.07	0.05	0.051	0.049	0.006	0.052	0.001	0.011	0.022	0.003	0.001	98.051
Weld Metal	0.52	12.6	0.12	1.12	0.009	0.062	0.19	0.21	0.01	0.015	0.009	0	0	85.132

Table 38.2 Chemical composition of the welding flux

SiO_2	Al_2O_2	MgO	CaF_2	CaO	MnO	TiO_2	Na_2O_3	FeO_3	C
15.82	19.23	28.40	22.15	7.10	2.21	1.48	0.97	2.12	0.51

Table 38.3 Process parameters

Polarity	DCEP
Electrode size (mm)	3
Electrode extension (mm)	15
Electrode feed rate (m/mm)	10
Base metal thickness (mm)	10
Current (Ampere)	90
Welding speed (mm/s)	6
Welding angle	30°
Shielding gas	95% Ar + 5% CO_2 (Argon + CO_2 are non-toxic)
Arc voltage (V)	18
Arc length (mm)	3

Table 38.4 Chemical composition of the fume

Element (% by weight)	Fe	K	Si	Mn	Zn	Ca	Na	Mg
	54.60	11.59	21.82	7.43	0.01	0.03	0.04	4.48

in Table 38.4. This analysis was done to determine the level of the toxic and poisonous elements and compounds that constitute the fumes.

From the constituent elements in Table 38.1, there was no trace of Potassium (K), therefore it is suspected that potassium silicate would have been used as the binder. However, since the iron, potassium, silicon, manganese and magnesium are present in large proportions as shown in Table 38.4, they are considered to be the only active elements in Table 38.4. The others are deemed to have an insignificant effect on fume toxicity.

The permissible exposure limit or occupational exposure limits (OELS) are the maximum permissible concentrations of a hazardous substance that most healthy adults may be repeatedly exposed to without suffering adverse health effects. These limits assume that the individual welders exposed to the substance is a healthy adult. OELs permit workers to be exposed to only very small quantities of these substances. OELs represent only minimum standards for the reason that

Table 38.5 osha exposure limits January 2007

Chromium (VI) (mg/m^3)	0.005
Manganese (fume) (mg/m^3)	0.2
Beryllium (mg/m^3)	0.002
Copper (mg/m^3)	0.2
Molybdenum (mg/m^3)	0.5
Nickel (mg/m^3)l	1.5
Vanadium oxide (fume) (mg/m^3)	0.05

measurements and workers' susceptibility is expected to vary [9]. Chan et al. [3] said that standards for welding fumes were established by the American Conference of Government Industrial Hygienists (ACGIH) in 1987, which defined a threshold limit value (TLV) of 5 mg/m^3 (8 h time weighted average). In 1989, the Occupational Safety and Health Administration (OSHA) also established an overall permissible exposure limit (PEL) of 5 mg/m^3, with much lower PELs for many of the particulate fume constituents. The quantities of the fume constituents of the permissible fume exposure limit set by the US Occupational Safety and Health Administration (OSHA) is shown in Table 38.5 [1].

38.2.1 Deleterious Effects of Welding Fumes and Gases

The deleterious effects of fumes, gases and organic vapors produced during any welding process are discussed hereunder taking into consideration the contributions of some investigators [8, 10–12].

It must be remembered that the welders are exposed to a cocktail of fumes and gases. It is therefore manifestly difficult to isolate and zero down the damaging effects of one element or compound to a specific injury.

Aluminum for instance, could cause lung damage, increases dementia, amyotropic lateral sclerosis, Parkinsons dementia, and alzheimers. Cadmium could cause serious pulmonary oedema, and chronic effects such as emphysema and kidney damage. Prolonged exposure to chromium may result in skin irritation, ulceration of the nasal septum, and a greater risk of lung cancer. Copper can cause respiratory irritation, nausea, acute lung damage and metal fume fever. Long-term exposure to fluorides may cause bone changes, and joint deterioration. Milder excessive exposure may have chronic effects such as pulmonary oedema, and skin rashes. Iron causes respiratory irritation, and is also capable of causing Siderosis, a benign accumulation of iron oxide in the lungs, leading to disturbances in lung function, as well as diabetes mellitus. Manganese could lead to manganism (manganese encephalopathy similar to Parkinson's disease). Symptoms are irritability, difficulty in walking, speech disturbances, compulsive behaviors and liver Cirrhosis. Lead exposure, could lead to peripheral neuropathy (damage that interrupts communication between the brain and other parts of the body and can impair muscle movement, prevent normal sensation in the extremities, and

cause pain), it also adversely affects the urinary, gastrointestinal, reproductive and skeletal systems. Molybdenum can cause respiratory irritation and impaired breathing, and Nickel could cause eye and throat irritation and is a known respiratory tract carcinogen. Nickel Carbonyl is extremely toxic on its own. Tin fumes are known to cause Stenosis, a benign pneumoconiosis. Vanadium could cause symptoms of eye and respiratory irritation, bronchitis, rhinitis, pulmonary oedema and pneumonia. Zinc is present in galvanized metals and could lead to metal fume fever. Beryllium could cause Berylliosis (scarring of the lungs preventing exchange of oxygen and carbon dioxide, and there is no known cure), as well as lung cancer. Cobalt when inhaled could cause Asthma, cumulative lung changes and dermatitis. Magnesium when inhaled, could cause metal fume fever (chills, fever, muscle aches).

As for gases, Ozone may be very detrimental to health, causing pulmonary congestion, oedema, and haemorrhage. Minute concentrations dry out the eyes and cause headaches. Prolonged exposure may result in severe changes in lung function. Oxides of nitrogen can cause eye, nose and lung irritation at 20–25 ppm. At higher concentrations, pulmonary oedema and other serious lung conditions can result. Carbon monoxide is absorbed into the bloodstream causing palpitations, headache, dizziness, confusion, and high concentrations may result in unconsciousness and eventual death. Hydrogen fluoride causes irritating to the eyes and respiratory tract. Overexposure can cause lung, bone, and kidney damage. Chronic exposure can result in chronic irritation of the nose, throat and bronchi. Oxygen deficiency occurs when welding in confined spaces, causing air displacement leading to dizziness, mental confusion, asphyxiation and eventual death.

Organic vapors produced during welding are aldehyde, phosgene, and phosphine. They often act as a severe irritants to the eyes, nose and the respiratory system and can also damage the kidneys and other organs.

38.2.2 Fume Formation Rate

The level of the welders' exposure to poisonous fumes determines the extent of their health risk assessment. Carter [13] said that the knowledge of fume emission rate is a key factor in estimating the risk of over-exposure. As earlier stated here, the OSHA set a permissible exposure limit to be less or equal to 5 mg/m^3, for eight working hours. At amounts above this limit, if these hazardous elements/particles are continuously inhaled for a considerable length of time, and particularly in a confined work space, the particles begin to accumulate in the lungs, and could cause irreversible injury. From micrographic examination, these fume particles, according to Stern [14], and Kalliomaki et al. [15], form long chains and rings and are therefore described as Polycrystalline, with an amorphous outer scale.

The methods adopted by Pires et al. [6] to obtain the Fume formation rate (FFR) were applied and this is expressed here under in Eq. 38.1

Table 38.6 Failure distribution time, reliability density and hazard rate in welding and fabrication firm

i	Time t (in yrs)	$\sum_{n=1}^{k} t$ (in yrs)	Reliability $R(t)$	Density $f(t)$	Hazard rate $\lambda(t)$
0	0.0	0.0	1.0	0.0137	0.0137
1	4.5	4.3	0.9412	0.0113	0.0120
2	5.2	9.5	0.8824	0.0113	0.0128
3	5.2	14.7	0.8235	0.0098	0.0119
4	6.0	20.7	0.7647	0.0075	0.0099
5	7.8	28.5	0.7059	0.0074	0.0104
6	8.0	36.5	0.6471	0.0072	0.0111
7	8.2	44.7	0.5882	0.0063	0.0108
8	9.3	54.0	0.5294	0.0049	0.0093
9	12.0	66.0	0.4706	0.0045	0.0095
10	13.2	79.2	0.4118	0.0044	0.0107
11	13.4	92.6	0.3529	0.0041	0.0117
12	14.3	106.9	0.2941	0.0039	0.0132
13	15.1	122.0	0.2353	0.0036	0.0154
14	16.2	138.2	0.1765	0.0030	0.0170
15	19.6	157.8	0.1176		

$$\text{FFR} = \frac{\text{Total weight of fumes produced}}{\text{Arc time}} \tag{38.1}$$

where FFR is expressed in terms of g/min. The arc time used for this study was 20 s, whereas the weight of fume collected was 0.065 g.

In this study the fume particle size, d_p was experimentally determined using

$$d_p = 5.5\ \text{FFR} \times \text{Total welding time} \tag{38.2}$$

The total welding time used was 120 s, where d_p is expressed in μm.

Carter [13] was of the opinion that the consumable with the highest emission rate will result in the highest exposure.

38.3 Hazard Profile of the Effect of Fume Constituents on Welders

The hazard profile of the fume constituents is a product of analyzing the meantime to failure (MTTF). This is the mean duration of time from a welder's initial exposure to welding fumes, to the time the welding related illnesses were first diagnosed. A summary of this investigation is shown in Table 38.6. The equations used for determining the reliability, failure density and hazard rate as shown in the tables can be found in Achebo and Oghoore [2].

In estimating the MTTF from the sample mean, Eq. 38.3 is applied,

$$\text{MTTF} = \sum_{i=1}^{n} \frac{t_i}{n} \tag{38.3}$$

A 95% confidence interval may be found from

$$\text{MTTF} = \frac{\left(\begin{array}{c} 0.0 + 4.3 + 9.5 + 14.7 + 20.7 + 28.5 + 36.5 + 44.7 + 54.0 \\ + 66.0 + 79.2 + 92.6 + 106.9 + 122.0 + 138.2 + 157.8 \end{array}\right)}{16}$$

$$= \frac{975.6}{16} = 60.98 \simeq 61$$

And the variance of the failure distribution, S^2 can be obtained from the sample variance as expressed in Eq. 38.4.

$$S^2 = \sum_{i=1}^{n} \frac{(t_i - \text{MTTF})^2}{n-1} \tag{38.4}$$

Alternatively, the variance can be expressed as

$$S^2 = \frac{\sum_{i=1}^{n} t_i^2 - n\text{MTTF}^2}{n-1} \tag{38.5}$$

S^2 is determined hereunder by substituting the values in the tables

$$S^2 = \frac{\left[\begin{array}{c} 4.3^2 + 9.5^2 + 14.7^2 + 20.7^2 + 28.5^2 + 36.5^2 + 44.7^2 + 54.0^2 + 66.0^2 \\ +79.2^2 + 92.6^2 + 106.9^2 + 122.0^2 + 138.2^2 + 157.8^2 - 16\left(61.0^2\right) \end{array}\right]}{15}$$

$$= \frac{97327 - 59536}{15} = \frac{37791}{15} = 2519.4$$

$$S = \sqrt{2519.4} = 50.2$$

From the t table, $t_{0.05,15} = 1.753$.

For large sample size of n failure times, an approximate 100% $(1 - \alpha)$ confident interval is applied to arrive at the MTTF. It is thus obtained as follows:

$$\text{MTTF} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}} \tag{38.6}$$

where n is the number of test trials.

From Eq. 38.6

$$61 \pm 1.753 \times \frac{50.2}{\sqrt{16}} = 61 \pm 22$$

$$= |39, 83| \text{ years}$$

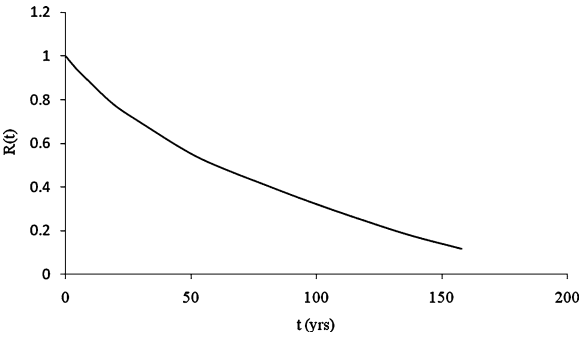


Fig. 38.1 The empirical reliability curve

From the tables, cumulative values corresponded to |8.1,13.3| years actual time. From the above derivation, under normal or safe operating conditions, the welders, to start with, are presumed to be in good health condition. Welders are expected to begin to experience the symptoms of any of the illnesses caused by welding fumes at the critical time frame of between 8.1 and 13.3 years of their service lives; considering the fume chemical composition, space ventilation, and respiratory control systems. Although, either acute or chronic infections or illnesses caused by welding particulates could occur at the critical time frame determined here, this; however, depends on the pre employment health condition of the welders.

It is therefore safe to conclude that welders that have these welding fume related illnesses earlier than the critical time frame may have been predisposed to health problems, probably due to possessing a diminished immune system, and possibly, a higher exposure rate. On the other hand, those welders that contracted these illnesses after the expected critical time may be deemed to have had better immune systems, still, they were affected by prolonged or accumulated inhaled welding fume particles in their lungs, predisposing them to chronic infections.

The relationship between the empirically derived or non-parametric reliability, $R(t)$, failure density function, $f(t)$, hazard rate, $\lambda(t)$ functions and their corresponding failure times were investigated as shown in Figs. 38.1, 38.2 and 38.3.

38.4 Discussion of Results

From the calculated fume formation rate, the average welder is exposed to 0.195 g/min compared to 0.84 and 3.1 g/min for SMAW and fluxed core arc welding (FCAW) electrodes, respectively, as obtained by other researchers [16], and also falls within the range of 0.02–0.28 g/min obtained by Pires et al. [6]. Whereas the mean agglomerated particle size of the fume was spherical in shape and coarse (Dp $>$ 1 μm) and was determined to be 2.14 μm. This value falls within 0.1–2.5 μm classified as fine particle size regimes in the fume particle size or number distribution tests conducted by Gonser et al. [16]. They also classified $<$0.1 μm as ultra fine particle size regimes.

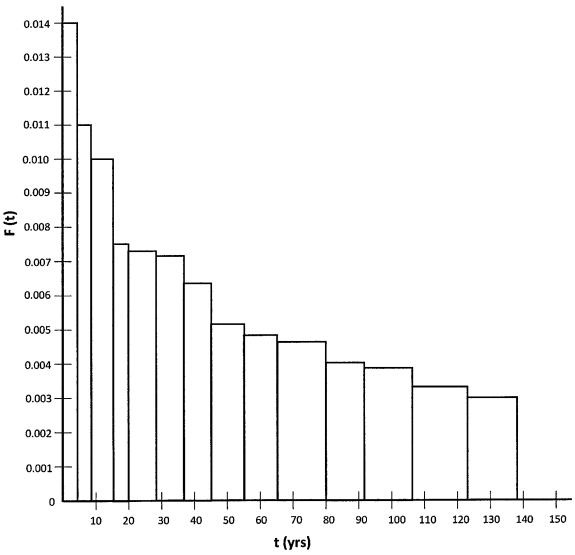


Fig. 38.2 The empirical failure density chart

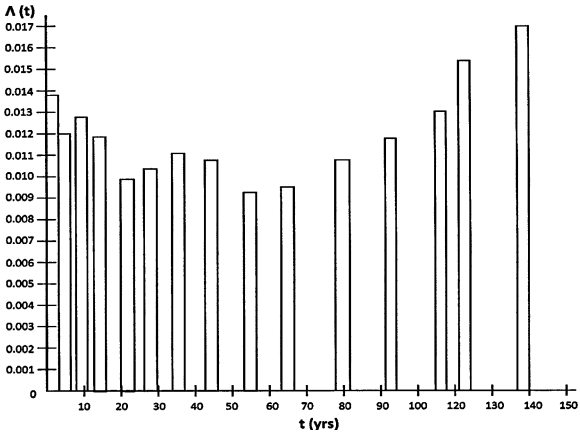


Fig. 38.3 The empirical hazard chart

Nederman [1] emphasized that the size of the particles is important because it controls the depth to which they penetrate the respiratory system. Particles larger than 5 μm are deposited in the upper respiratory tract. Particles in the range of 0.1–5 μm which includes welding fumes, penetrate the inner parts of the lungs (the alveoli) and are deposited there. Over time, the particles can even reach the bloodstream. However, particles between 0.1 and 5 μm are usually very toxic and could lead to immediate death.

From the constituent elements present in the chemical composition of the welding fume, it can be seen that the dominant constituent elements are Fe(iron), K(potassium), Si(silicon), Mn(manganese), and Mg(magnesium). Therefore, the welders were invariably exposed to these constituents. From epidemiological studies,

the possible adverse health effect associated with these elements can be found in session three.

The analysis done by employing the probability failure functions, used by Achebo and Oghoore [2] reveal that welders are expected to have fume related illnesses within 8.1 and 13.3 years from the date of their employment. From the analysis done, Fig. 38.1 shows that the increase in the mean time to failure correspondingly reduces reliability of the efficiency of the welding process. This means that the less there are welders affected by the consequences of welding fumes, the healthier and more productive they become and the more reliable the method of controlling the incidence of fume exposure within the critical time frame of 8.1–13.3 years. Figure 38.2 explains the fact that the higher the mean time to failure the lower the failure density distribution. Figure 38.2 further explains the theory of conditional density function, which is the probability that failure could take place at a time before the critical time frame. In this case, it explains that some illnesses that would have ordinarily been expected to occur before the critical time frame, were found to occur after the critical time frame. Therefore as the welders avoid the over exposure to welding fumes and gases, the expected time for the welder to become sick is prolonged and extended. Hazard rate could be seen as the instantaneous failure rate or illnesses caused as a result of the accumulated fume particle present in the welders' lungs, it could either be in the form of acute or chronic illnesses. The turbulent pattern of flow of the bar charts arrangement in Fig. 38.3, that is from high to low, then high again, shows that the hazardous nature of the fume is not proportional or homogeneous. Figure 38.3 clearly indicates that the hazards caused by these fumes and gases, to a great extent depend on the state of health of the welders. The hazard profile spans its active time from before the critical time, $<t$, to the critical time frame, t, then to the chronic stage, $\Delta t + t$.

Recognizing and identifying the associated ill health effects as well as the quantity of the harmful constituents that have made their way to the lungs of the workers, can help the filler wire/coated electrode manufacturers to reduce these elements to the barest minimum at source. This would reduce the incidence or occurrence of these illnesses. Pires et al. [6] said that fume control at source, by modification of procedures and/or consumables can be used to complement the existing control strategies and this, they said would provide a healthier environment for the welders. This data is expected to help management to control and make policies that would provide a healthier environment for welders.

38.5 Conclusion

The effect of welding fumes and associated toxic gases on 15 welders working in a well established welding and fabrication firm in Nigeria was investigated. The welding process was carried out using selected welding process parameters. The chemical composition of the fumes evolved from the welding process revealed

the constituent elements that were expected to have been inhaled by the welders and the associated health hazards have been discussed, applying the probability failure functions model used by Achebo and Oghoore [2]. This study agrees with other investigators that fume control should first of all, be addressed at source, complimenting the other control strategies. From the findings of this study, it is suggested that the firm's management should provide a well ventilated welding environment; provide respiratory equipment for its welders; ensure that welders go for routine clinical checks, and filler wires or coated electrodes whose constituent elements conform to the occupational exposure limits (OEL) should be utilized, to protect the welders from the danger posed by potentially harmful constituents. The aim of this study has been successfully achieved in terms of determining the likely causes of the fume related illnesses and the asymptomatic hazards experienced by 15 selected welders, and cogent suggestions for bringing about healthier and more efficient welding conditions, devised.

References

1. Nederman (2010) Welding fume hazards and prevention. http://www.nederman.com.br/pdf/fumos_gases_solda.pdf
2. Achebo JI, Oghoore O (2010) A nonparametric analysis of asymptomatic hazard rates in a brewing plant using the probability failure functions. Lecture notes in engineering and computer science. Proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, pp 2381–2384
3. Chan W, Gunter KL, Sutherland JW (2002) An experimental study of the fume particulate produced by the shielded metal arc welding process. http://www.me.mtu.edu/~jwsuther/Publications/72_Chan_Gunter_Sutherland_MS29_year.pdf
4. Stern RM, Chemical A (1979) Physical and biological assay of welding fume. The Danish Welding Institute, Copenhagen, p 2
5. Hilton DE, Plumridge PN (1991) Particulate fume generation during GMAW and GTAW. Weld Metal Fabr 62:555–560
6. Pires I, Quintino L, Miranda RM, Gomes JFP (2006) Fume emissions during gas metal Arc welding. Toxic Environ Chem 88(3):385–394
7. Lyttle K (2004) Optimizing consumable selection increases productivity, decreases fumes. Gas Weld Distr 45–47
8. Spear JE (2004) Welding fume and gas exposure. J.E. Spear consulting, LLC, 19314 Timber Ridge Drive, Suite 100 Magnolia, Texas 77355
9. Government of Alberta employment and immigration welder's guide to the hazards of welding gases and fumes (2009). Workplace Health Safety Bulletin. http://employment.alberta.ca/documents/WHS/WHS-PUB_ch032.pdf
10. Government of South Australia welding metal fumes (2000). http://www.safework.sa.gov.au/uploaded_files/gs33i.pdf
11. Zenith Insurance Company, Health hazards of welding (2009). http://www.thezenith.com/employers/services/pi/indsaf/agr/rmb/agriculture_welding_rm129.pdf
12. Bailey S, Clark M, Welding fumes hazards and control. http://www.agccolorado.org/site/publisher/files/safety/welding_fume_controls.ppt
13. Carter GJ (2005) Laboratory methods for measuring fume and gas emission rates and sampling for analysis paper presented at the international conference on health and safety in welding and allied processes. Kopenhagen, Denmark, 9–11 May

14. Stern R (1976) Production and characterization of a reference standard welding fume, Rapport SVC, Kopenhagen
15. Kalliomaki P, Sutinen S, Kelha V, Lakomaa E, Sortti V, Sutinen S (1979) Amount and distribution of fume contaminants in the lungs of an arc welder post mortem. Brit J Ind Med 36:224–230
16. Gonser MJ, Lippold JC, Dickinson DW, Sowards JW, Ramirez AJ (2010) Characterization of welding fume generated by High–Mn consumables. Weld J 89(2):25s–33s

Chapter 39
Classification and Measurement of Efficiency and Congestion of Supply Chains

Mithun J. Sharma and Song Jin Yu

Abstract In order to enhance and extend the variation of date envelopment analysis (DEA) methodology, this chapter serves to supplement the DEA literature in its application to supply chain efficiency measurement. In addition an examination of input congestion is carried out indicates that a managerial inefficiency exist in the different process cycles of supply chains. However, presence of congestion indicates the inability to dispose of unwanted inputs without incurring cost. Using the DEA variation, supply chains are partitioned into three levels/stratums namely 'best-in-class', 'average' and 'laggard'. Substantial performance inefficiency is uncovered in the four process cycle dimensions. Relatively, down-stream process cycles of the supply chain exhibit better performance than the up-stream process cycles. Our innovative approach identifies areas for improved supply chain performance over the four process cycles. The classification of supply chains serve as a guideline for best practices, and projects directly to the best-in-class.

39.1 Introduction

Supply chain is a combined system which comprises planning, sourcing, making and development of processes with its constituent parts to include material suppliers, production facilities, distribution centers and customers forward flow of

M. J. Sharma (✉)
Centre for Management Studies, Dibrugarh University, Dibrugarh, Assam 786004, India
e-mail: mithunjsharma@dibru.ac.in

S. J. Yu
Department of Shipping Management, Korea Maritime University, Busan 609781, South Korea
e-mail: coppers@hhu.ac

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_39,
© Springer Science+Business Media B.V. 2011

material as well as feedback flow of information [1]. A company can identify its supply chain by first selecting a particular product group or product family. Then it should trace the flow of materials and information from the final customer backward through the distribution system, to the manufacturer and then to the suppliers and the source of raw material. The entire chain of activities and processes is known as the supply chain for that product group. An increasingly popular perspective today is to view the flow of materials from the point of conception to consumption as a system which involves strategic coordination of each echelon to be managed. This perspective is commonly referred to as supply chain management.

To improve efficiency and effectiveness of a supply chain it is crucial to increase co-ordination both across firms and within firms which are members of the supply chain. A typical firm consists of separate departments which manages the different aspects of the supply chain. For instance, purchasing takes care of the suppliers and raw materials inventory, operations takes care of manufacturing and work-in-process inventory and marketing manages demand and finished goods inventory. When these departments lack coordination, there are dramatic effects on supply chain within the firm as well as outside the firm. Thus measuring supply chain performance is the first step towards improvement.

Performance measurement plays an essential role in evaluating production because it can define not only the current state of the system but also its future. According to Dyson [2] performance measurement helps move the system in the desired direction through the effect exerted by the behavioral responses towards these performance measures that exist within the system. Mis-specified performance measures, however, will cause unintended consequences, with the system moving in the wrong direction [2]. The underlying assumption behind this claim is the role or presence of drivers such as efficiency and effectiveness in the composition of performance. To put it in a simple way, efficiency in Dyson's [2] claim is 'doing things right' and effectiveness is 'doing the right thing'. The combination of these two key drivers helps move the system in the right direction by doing the right thing.

The efficiency is determined by using a variation of frontier estimation especially data envelopment analysis (DEA) amidst multiple inputs and outputs. In particular, DEA methodology has proved to be powerful for benchmarking and identifying efficient frontiers especially for single producers or decision making units (DMU). Literature reviews, such as the excellent bibliography in Seiford [3], reveal that research examining the use of mathematical programming and associated statistical techniques to aid decision-making in supply chain benchmarking is lacking. Liang et al. [4] points out that traditionally most models (deterministic and stochastic) dealt with isolated parts of supply chain systems. Liang et al. [4] developed a Stackelberg co-operative model to evaluate the efficiency of SC members using DEA but their study was neither empirical nor showed any relationship of co-operation among members. An empirical study to evaluate the efficiency of whole supply chain was done by Reiner and Hoffman [5]. They tried to evaluate the processes in a supply chain using the performance measure of

SCOR [6]; however they considered various processes of a single supply chain instead of multiple chains. This leaves us with a literature gap and a question on how to measure the performance of supply chain considering each supply chain as meta-DMU. Research is required to find out how to measure the efficiency of a supply chain keeping an eye on key performance metrics that can cover all the interfaces in a supply chain.

Some researchers have tried to evaluate the chain in a serial order while others have tried to use a single performance measure [7]. Some others like Chen and Zhu [8] have provided two approaches in modeling efficiency as a two-stage process. Golany et al. [9] provided an efficiency measurement framework for systems composed of two subsystems arranged in series that simultaneously compute the efficiency of the aggregate system and each subsystem. Zhu [10], on the other hand, presented a DEA-based supply chain model to define and measure the efficiency of a supply chain and that of its members. Fare and Grosskopf [11] and Castelli et al. [12] introduced the network DEA model, in which the interior structure of production units can be explicitly modeled.

However, a supply chain is a sequence of processes and flows that take place within and between different stages and combine to fill a customer need for a product. The objective of every supply chain is to maximize the overall value generated. The value a supply chain generates is the difference between what the final product is worth to the customer and the effort the supply chain expends in filling the customer's request. For most commercial supply chains, value will be strongly correlated with *supply chain profitability*. *Supply chain profitability* is the total profit shared across all the supply chain stages [13]. All the above mentioned studies evaluated supply chain in stages ignoring the essence of processes that knits the stages of supply chain. By focusing on the process as the unit of analysis, the management of inter-organizational relations in a way which is generally known as network, on performance is analyzed.

In order to enhance and extend the variation of DEA methodology this chapter serves to supplement the DEA literature in its application to supply chain. This chapter extends the work published in WCE [14] and proposes a model to evaluate the overall supply chain efficiency. Using the DEA variation of Sharma and Yu [15], the supply chains are partitioned into three levels/stratums namely 'best-in-class', 'average' and 'laggard' [16]. In addition, an examination of input congestion is carried out indicates that a managerial inefficiency [17] exist in 'average' and 'laggard' supply chains. By simply reconfiguring these excess resources it may be possible to increase output without reducing the inputs. Equipped with this knowledge, managers will be better able to determine when large reengineering projects are necessary versus minor adjustments to existing business processes. The object oriented DEA models to classify and measure efficiency and congestion of supply chains are discussed in Sect. 39.2, followed by identifying inputs and outputs of each process cycle of supply chain and the resulting empirical findings in Sect. 39.3. Finally, Sect. 39.4 summarizes and concludes the chapter.

39.2 DEA Supply Chain Models

39.2.1 DEA Models to Measure Supply Chain Efficiency

There are some issues related to measuring the efficiency of a supply chain using DEA. The first is that supply chain operations involve multiple inputs and outputs of different forms at different stages and second is that the performance evaluation and improvement actions should be coordinated across all levels of production in a supply network. In this chapter, we evaluate supply chain stages in process cycles keeping the essence of processes that knits the stages of supply chain. By focusing on the process as the unit of analysis, the management of inter-organizational relations in a way which is generally known as network, on performance will be analyzed.

DEA models are classified with respect to the type of envelopment surface, the efficiency measurement and the orientation (input or output). There are two basic types of envelopment surfaces in DEA known as constant returns-to-scale (CRS) and variable returns-to-scale (VRS) surfaces. Each model makes implicit assumptions concerning returns-to-scale associated with each type of surface. Charnes et al. [18] introduced the CCR or CRS model that assumes that the increase of outputs is proportional to the increase of inputs at any scale of operation. Banker et al. [19] introduced the BCC or VRS model allowing the production technology to exhibit increasing returns-to-scale (IRS) and decreasing returns-to-scale (DRS) as well as CRS.

A common approach to evaluate supply chain in particular two stage DEA is that the first stage uses inputs to generate outputs that then become the inputs to the second stage. The second stage thus utilizes these first stage outputs to produce its own outputs under CRS and VRS assumptions [20, 21]. However, a supply chain is a sequence of processes and flows that take place within and between different stages and combine to fill a customer need for a product with an objective to maximize the overall value of the supply chain. Previous studies evaluated supply chain in stages ignoring the essence of processes that knits the stages of supply chain. Therefore evaluating supply chain processes and sub processes will help to effectively analyze supply chain as a whole. A detailed description of supply chain processes along with inputs and outputs of each process cycles are discussed in [14].

39.2.1.1 The BCC Supply Chain Model

The input-oriented BCC model evaluates the efficiency of DMU_o ($o = 1, \ldots, n$) by solving the following envelopment form: (BCC_o)

$$\mathrm{Min}_{\theta_B, \lambda} \theta_B$$

Subject to

$$\theta_B x_o - X\lambda \geq 0$$
$$Y\lambda \geq 0$$
$$e\lambda \geq 0$$

where θ_B is a scalar.

The dual multiplier form of this linear program (BCC_o) is expressed as

$$\text{Max}_{v,u,u_o} z = uy_o - u_o$$

Subject to

$$vx_o = 1$$
$$-vX + uY - u_o e \leq 0$$

$v \geq 0,\ u \geq 0$ and u_o free in sign, where v and u are vectors and z and $u_o e$ are scalars and the latter, being 'free in sign', may be positive or negative or zero. The equivalent BCC fractional program is obtained from the dual program as

$$\text{Max}\frac{uy_o - u_o}{vx_o}$$

Subject to

$$\frac{uy_j - u_o}{vx_j} \leq 1 (j = 1, \ldots, n)$$

$v \geq 0,\ u \geq 0$, and u_o free.

The primal problem (BCC_o) is solved using two-phase procedure. In the first phase, we minimize θ_B and, in the second phase, we minimize the sum of the input excesses and output shortfalls, keeping $\theta_B = \theta_B^*$. An optimal solution for (BCC_o) is represented by $\theta_B^*, \lambda^*, s^{-*}, s^{+*}$, where s^{-*} and s^{+*} represent the maximal input excesses and output shortfalls, respectively.

BCC-Efficiency. If an optimal solution $\theta_B^*, \lambda^*, s^{-*}, s^{+*}$ obtained in this two phase process for (BCC_o) satisfies

$$\theta_B = 1$$

And has no slacks ($s^{-*} = 0$ and $s^{+*} = 0$), then the DMU_o, we define its reference set, E_o, based on an optimal solution λ^* by

$$E_o = \{j | \lambda_j^* \geq 0\} (j \in \{1, \ldots, n\})$$

If there are multiple optimal solutions, we can choose any one to find that

$$\theta_B^* x_o = \sum_{j \in E_o} \lambda_j^* x_j + s^{-*}$$

$$y_o = \sum_{j \in E_o} \lambda_j^* y_j - s^{+*}$$

Thus the improvement path via the BCC projection,

$$\hat{x}_o \Leftarrow \theta_B^* x_o - s^{+*}; \quad \hat{y}_o \Leftarrow y_o + s^{+*}$$

The above VRS setting is proposed for a single process of supply chain. In Kao and Hwang [20] model to measure the two-stage processes, they combine the processes in a multiplicative (geometric) manner. In our proposed VRS model we combine the processes in an arithmetic mean approach since the processes are interlinked and not independent. The same averaging method is applied to CCR model.

39.2.1.2 Input Congestion Supply Chain Model

Congestion is said to occur when the output that is maximally possible can be increased by reducing one or more inputs without improving any other input or output. Conversely, congestion is said to occur when some of the outputs that are maximally possible are reduced by increasing one or more inputs without improving any other input or output. For example, excess inventory cluttering a factory floor in a way that interferes with production. By simply reconfiguring this excess inventory it may be possible to increase output without reducing inventory. This improvement represents the elimination of inefficiency that is caused by the way excess inventory is managed. There are many models dealing with congestion but we start with FGL (Fare et al. [22, 23]) because it has been the longest standing and most used approach to congestion in the DEA literature. Fare, Grosskopf and Lovell (FGL) approach proceeds in two stages. The first stage uses an input oriented model as follows (Fare et al. [23]):

$$\theta^* = \min \theta$$

Subject to

$$\theta x_{io} \leq \sum_{j=1}^{n} x_{ij}\lambda_j, \quad i = 1, 2, \ldots, m \tag{39.1}$$

$$y_{ro} \leq \sum_{j=1}^{n} y_{rj}\lambda_j, \quad r = 1, 2, \ldots, s$$

$$\lambda_j \geq 0, \quad j = 1, 2, \ldots, n$$

where $j = 1, 2, \ldots, n$ indexes the set to DMUs (decision making units) which are of interest. Here is the observed amount of input $i = 1, 2, \ldots, m$ used by DMU_j and y_{ro} is the observed amount of output $r = 1, 2, \ldots, s$ associated with DMU_o is the $\text{DMU}_j = \text{DMU}_o$ to be evaluated relative to all DMU_j (including itself).

The objective is to minimize all the inputs DMU_o in the proportion θ^* where, because the $x_{io} = x_{ij}$ and $y_{ro} = y_{rj}$ for $\text{DMU}_j = \text{DMU}_o$ appear on both sides of the constraints, the optimal $\theta = \theta^*$ does not exceed unity and the non-negativity of the λ_j, x_{ij}, and y_{rj} implies that the value of θ^* will not be negative under the optimization in (39.1). Hence,

$$0 \leq \min \theta = \theta^* \leq 1 \tag{39.2}$$

We now have the following definition of technical efficiency and inefficiency. Technical efficiency is achieved by DMU_o if and only if $\theta^* = 1$. Technical inefficiency is present in the performance of DMU_o if and only if $0 \leq \theta^* \leq 1$. Next, FGL then go on to the following second stage model,

$$\beta^* = \min \beta$$

Subject to

$$x_{io} = \sum_{j=1}^{n} x_{ij}\lambda_j, \quad i = 1, 2, \ldots, m \tag{39.3}$$

$$y_{ro} \leq \sum_{j=1}^{n} y_{rj}\lambda_j, \quad r = 1, 2, \ldots, s$$

$$\lambda_j \geq 0, \quad j = 1, 2, \ldots, n$$

Note that the first $i = 1, 2, \ldots, m$ inequalities in (39.1) are replaced by Eq. 39.3. Thus slack is not possible in the inputs. The fact that only the output can yield non-zero slack is then referred to as weak disposal by Fare et al. [22]. Hence, we have $0 = \theta^* \leq \beta^*$. FGL use this property to develop a measure of congestion,

$$0 \leq C(\theta^*, \beta^*) = \frac{\theta^*}{\beta^*} \leq 1 \tag{39.4}$$

Combining model (39.1) and (39.3) in a two-stage manner, FGL utilizes this measure to identify congestion in terms of the following conditions,

1. Congestion is identified as present in the performance of DMU_o if and only if

$$C(\theta^*, \beta^*) < 1 \tag{39.5}$$

2. Congestion is identified as not present in the performance of DMU_o if and only if

$$C(\theta^*, \beta^*) = 1$$

Our proposed congestion model will use arithmetic mean of the congestion scores of each process cycle to check the presence or absence of congestion in the overall supply chain θ^{o*}/β^{o*}.

39.2.1.3 Supply Chain Classification Model

Classification of supply chain is required to standardize the sets of efficient and inefficient supply chains for step-wise improvement, otherwise not possible with the traditional DEA. To classify the set of supply chains, we modify the algorithm developed by Sharma and Yu [15] to segment the supply chains into three classes namely, best-in-class, average, and laggard chains. The modified algorithm is as follows:

Assume there are DMUs, each with m inputs and s outputs. We define the set of all DMUs as J^1, $J^1 = \text{DMU}_j$, $j = 1, 2, \ldots, n$ and the set of efficient DMUs in J^1as E^1. Then the sequences of J^1 and E^1 are defined interactively as $J^{l+1} = J^l - E^l$ where $E^l = \text{DMU}_p \in J^l | \phi_p^l = l$, and ϕ_p^l is optimal value to the following linear programming problem

$$max_{\lambda_i, \phi} \phi_p^l = \phi$$

Subject to

$$\sum_{i \in F(j^l)} \lambda_i x_{ji} - x_{jp} \leq 0 \quad \forall j$$

$$\sum_{i \in F(j^l)} \lambda_i y_{ki} - \phi y_{kp} \geq 0 \quad \forall k$$

$$\lambda_i \geq 0,\ i \in F\left(j^l\right)$$

where $k = 1$ to s, $j = 1$ to m, $i = 1$ to n, y_{ki} = amount of output k produced by DMU_i; x_{jp} = input vector of DMU_p, x_{ji} = amount of input j utilized by DMU_i; y_{kp} = output vector of DMU_p.$i \in F(j^l)$ in other words $\text{DMU}_i \in j^l$, i.e. $F(\cdot)$ represents the correspondence from a DMU set to the corresponding subscript index set. The following algorithm accomplishes subsequent stratum.

Step 1 Set $l = 1$. Evaluate the set of DMUs, j^l, to obtain the set E^1, of the first level frontier DMUs (which is equivalent to classical CCR DEA model), i.e. when $l = 1$, the procedure runs a complete envelopment model on all n DMUs and E^1 consists of all of the DMUs on the resulting overall best-practice efficient frontier.

Step 2 Exclude the frontier DMUs from future DEA runs and set $J^{(l+1)} = J^l - E^1$.

Step 3 If $J^{l+1} = 3E^{l+1}$, then stop. Otherwise, evaluate the remaining subset of inefficient DMUs, $J^{(l+1)}$, to obtain the new best-practice frontier E^{l+1}.

Stopping rule The algorithm stops when $J^{l+1} = 3E^{l+1}$.

39.3 Model Application

As already mentioned at the outset, a supply chain is a sequence of processes and flows that take place within and between different stages and combine to fill a customer need for a product with an objective to maximize the overall value of the

supply chain. Previous studies evaluated supply chain in stages ignoring the essence of processes that knits the stages of supply chain. Therefore evaluating supply chain processes and sub-processes will help to effectively analyze supply chain as a whole. Davenport and Short [24] define 'processes a set of logically related tasks performed to achieve a defined business outcome and suggest that processes can be divided into those that are operationally oriented (those related to the product and customer) and management oriented (those that deal with obtaining and coordinating resources).

The processes of a supply chain are divided into a series of cycles, each performed at the interface between two successive stages of a supply chain. A cycle view of the supply chain clearly defines the processes involved and the owners of each process. This view is very useful when considering operational decisions because it specifies the roles and responsibilities of each member of the supply chain and the desired outcome of each member of the supply chain and the desired outcome for each process. To evaluate the efficiency of supply chain we will consider four process cycles namely—customer order cycle, manufacturing process cycle, replenishment process cycle, and procurement process cycle. A detailed description of these cycles is provided by Sharma and Yu [14]. The inputs and outputs of each of these cycles taken from Sharma and Yu's studies [14].

39.3.1 BCC Supply Chain Model Application

The efficiency results of the CCR and BCC model for 11 supply chain sub-processes of a particular product. First, the efficient supply chains, in each process cycle are: customer order cycle (1, 4, 7, 9, and 11) replenishment process cycle (1, 2, 5, 6, 8, 11), manufacturing process cycle (2, 4, 6) and procurement process cycle (5, 6, 9). The RTS efficiency score is calculated as the ratio of CCR efficiency score to BCC efficiency score. The customer order cycle, the BCC efficient but not scale-efficient process, cycles were operating on an increasing returns to scale (IRS) frontier. For customer order cycle, five BCC-efficient retail chains were operating on IRS and four on decreasing returns to scale (DRS) frontiers. Of the BCC-inefficient supply chains, 64 and 20% were in the IRS region in cycles 1 and 2, respectively. As economists have long recognized, an IRS frontier firm would generally be in a more favorable position for expansion, compared to a firm operating in a constant returns to scale (CRS) or DRS region. Note that the concept of RTS may be ambiguous unless a process cycle is on the BCC-efficient frontier, since we classified RTS for inefficient process cycles by their input oriented BCC projections. Thus, a different RTS classification may be obtained for a different orientation, since the input-oriented and the output-oriented BCC models can yield different projection points on the VRS frontier. Thus, it is necessary to explore the robustness of the RTS classification under the output oriented DEA method. Note that an IRS DMU (under the output-oriented DEA method) must be termed as IRS by the input oriented DEA method. Therefore, one only needs to check the CRS

Table 39.1 Input congestion results of process cycles and overall supply chain

DMU ID	Customer order cycle			Replenishment cycle			Manufacturing cycle			Procurement cycle			Overall congestion
	θ^*	β^*	$\frac{\theta^*}{\beta^*}$	θ^*	β^*	$\frac{\theta^*}{\beta^*}$	θ^*	β^*	$\frac{\theta^*}{\beta^*}$	θ^*	β^*	$\frac{\theta^*}{\beta^*}$	$\frac{\theta^{o*}}{\beta^{o*}}$
1	1.0	1.0	1.0	1.0	1.0	1.0	0.2	0.2	1.0	0.9	0.9	1.0	1.00
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.00
3	0.4	0.4	0.9	0.8	0.8	0.9	0.0	0.0	0.6	0.1	0.1	0.9	0.87
4	1.0	1.0	1.0	0.6	0.6	1.0	0.8	1.0	0.8	0.9	0.9	1.0	0.96
5	0.5	0.5	1.0	1.0	1.0	1.0	0.3	0.3	0.7	0.6	1.0	0.6	0.86
6	0.6	0.6	0.9	0.6	1.0	0.6	0.7	1.0	0.7	1.0	1.0	1.0	0.84
7	0.9	1.0	0.9	0.6	0.7	0.9	0.0	0.0	0.6	0.1	0.1	1.0	0.90
8	0.5	0.5	0.9	1.0	1.0	1.0	0.1	0.1	1.0	0.2	0.3	0.9	0.97
9	1.0	1.0	1.0	0.9	0.9	1.0	0.7	0.7	1.0	1.0	1.0	1.0	1.00
10	0.7	0.7	1.0	1.0	1.0	1.0	0.1	0.1	1.0	0.1	0.1	0.9	0.97
11	0.7	0.7	1.0	0.9	1.0	0.9	0.1	0.1	1.0	0.1	0.1	1.0	0.99

and DRS supply chain processes in the current study. Using the input oriented approach, we discover only two DRS supply chain processes in replenishment cycle (DMUs 2, 4, 6 and 9) and seven DRS (DMUs 1, 3–7 and 9) in the manufacturing cycle. These results indicate that (i) in general; the RTS classification under different process cycle is independent of the orientation of DEA model; and (ii) there are serious input deficiencies in manufacturing cycle at the current usage quantities derived from engineering and process design. The overall supply chain of one chain i.e. DMU 2 is found to be efficient in both the CCR and BCC setting. However, there is one chain found efficient in BCC setting i.e. DMU 6.

39.3.2 Input Congestion Supply Chain Model Application

In Table 39.1, we focus on the points for DMUs 3, 6, 7, and 8, of customer order cycle which are the only ones that satisfy the condition for congestion specified in Eq. 39.5. For DMUs 3, 6, 7, and 8 in the Table 39.1 and coupling this value we obtain congestion efficiency as 0.91, 0.96, 0.97 and 0.98, respectively. Around 36% of the supply chains have exhibited input congestion under VRS technologies. The inputs technological functionality and sales order by FTE in a VRS technology shows the congestion of sales order by FTE is 18.36% of the corresponding technological functionality input level.

Around 45% of the supply chains have exhibited input congestion under VRS technologies in the replenishment process cycle. In the replenishment process cycle we focus on DMUs 2, 3, 6, 7, and 11. We obtain congestion efficiencies of 0.98, 0.97, 0.66, 0.98 and 0.96 for these supply chains. The inputs technological functionality and sales order by FTE same as the customer order cycle, in a VRS technology shows the congestion of sales order by FTE is 26.66% of the corresponding technological functionality input level.

In the manufacturing cycle, the focus DMU points are 3–7 which are the ones that satisfy the conditions of congestion specified in Eq. 39.5. The congestion efficiencies for these DMUs are 0.66, 0.87, 0.79, 0.77, and 0.66, respectively. The inputs bill-of-materials (BOM), usage quantity, independent demand ratio shows congestion by 15.2, 22.4, and 2.5%, respectively. The residual score in manufacturing cycle largely indicates the scope for efficiency improvements resulting from less efficient work practices and poor management, but also reflect differences between operating environments in these five supply chains.

The DMUs 2, 3, 5, 8 and 10 of the procurement cycle exhibits the presence of congestion. The congestion efficiency for these supply chains is found to be 0.94, 0.94, 0.66, 0.93, and 0.90, respectively. The inputs purchased item shows congestion by 23.3% of the corresponding input direct material cost.

Starting with input (in the form of technological functionality, order by FTE, BOM, usage quantity, independent demand ration, purchased items, direct material cost) at $x = 0$ the output, y_o, measured in fill rate, cycle inventory, inventory replenishment cycle time, finished product cycle time, end time, on time ship rate and DSA, can be increased at an increasing rate until x_o is reached at output y_o. This can occur, for instance, because an increase in the technological functionality, usage quantity, and purchased items makes it possible to perform tasks in a manner that would not be possible with a smaller number of inputs. From x_o to x_1 however, total output continues to increase, but at a decreasing rate, until the maximum possible output is reached at y_1. Using more input results in a decrease from this maximum so that at x_2 we have $y_2 < y_1$ and $y_1 - y_2$ is the amount of output lost due to congestion. Under congestion, the inability to dispose of unwanted inputs increases costs.

The overall supply chain congestion θ^{o*}/β^{o*} with 1 indicates absence of any congestion found in supply chains 1, 2, and 9. The rest of the supply chain exhibits at least some amount of congestion. Although in a few chains the congestion is negligible for instance the DMUs 8, 10, and 11. The highest amount of congestion is found in DMU 6 followed by DMU 5.

39.3.3 Supply Chain Classification Model Application

We analyzed the aggregated metrics of the companies using the modified algorithm of Chen et al. [21] to determine whether their performance ranked as best-in-class (36%), average (27%), or laggard (37%). In addition to having common performance levels, each class also shared characteristics in four process cycles: (1) customer order cycle (balances customer demand with supply from manufacturers); (2) replenishment process cycle (Balances retailer demand with distributor fill rate); (3) manufacturing cycle (balances the percentage mix of demand for an item from independent (outside sources) versus dependent (inside sources) across all supply chain stages); (3) procurement process cycle (balances delivery schedule adherence (DSA) for the timeliness of deliveries from suppliers).

Table 39.2 Results of classified supply chains

	Best-in-class (E^1)	Average (E^2)	Laggard (E^3)
Classes of efficient supply chains (%)	36	27	37
Customer order cycle (%)			
Balances customer demand with supply from manufacturers	66	53	48
Replenishment process cycle (%)			
Balances retailer demand with distributor fill rate	55	31	23
Manufacturing process cycle (%)			
Balances the percentage mix of demand for an item from independent (outside sources) vs. dependent (inside sources) across all supply chain stages	65	44	36
Procurement process cycle (%)			
Balances delivery schedule adherence (DSA) for the timeliness of deliveries from suppliers	52	48	45

The characteristics of these performance metrics serve as guideline for best practices, and correlate directly with best-in-class performance.

Based on the findings in Table 39.2 derived from the context dependent DEA algorithm (modified), the best-in-class supply chains reveal the optimal utilization of technological functionality along with the use of state-of-art technology. The average and laggard supply chains on the other hand must upgrade their technological functionality towards fast, responsive, and structured supply chains where customer responsiveness and collaboration are necessary ingredients for continued and relentless inventory, margin, working capital, and perfect order-related success.

Best-in-class supply chains processes sales order by full time employees 24–32% more than the average and laggard chain in the replenishment process cycle. As well as the fill rate and the time required to deploy the product to the appropriate distribution center is 28% higher than the average and laggard supply chains.

In the manufacturing cycle front the inventory optimization goals are well served by best-in-class chains. They work closely with their trading partners, including suppliers, distributor, and retailers to reduce the pressure of increased lead times and potentially lower inventory levels for the chain. Due to this close collaboration, the finished product cycle time (average time associated with analyzing activities, such as: package, stock, etc.) and end item (the final product sold to a customer) less relative to average and laggard supply chains by 34.5%.

On time ship rate (percent of orders where shipped on or before the requested ship date) and delivery schedule adherence (DSA) (a business metric used to calculate the timeliness of deliveries from suppliers) in the procurement cycle does

not show any significant difference among the best-in-class, average and laggard supply chains. There is only a 5% difference in the performance of this supplier manufacturer interface.

39.4 Conclusion

This chapter analyzes the process cycles of 11 supply chains using an innovative DEA model. Close to 45% of the supply chains were inefficient in four process cycles namely—customer order cycle, replenishment process cycle, manufacturing cycle and procurement cycle. Further, most supply chains exhibited DRS in manufacturing cycle and procurement cycle, while some of them exhibited IRS in customer order cycle and replenishment process cycle. This suggests that up-stream components of the supply chain may have a negative effect on finished product cycle time and end item. Having examined performance at process cycle of a supply chain, the current study employs a procedure by FGL [12] with modification to identify the presence of congestion in the chains that may hinder improvement projection of the inefficient chains incurring some cost. Then a context-dependent DEA model is used to classify the chains into three categories—best-in-class, average, and laggard chains. The characteristics of these performances metrics serve as guideline for best practices, and correlate directly with best-in-class performance. Finally, our examination of supply chain data set indicates that the gap in performance is higher in the down-stream relative to up-stream.

Acknowledgments This research was a part of the results of the project called "*International Exchange & Cooperation Project for Shipping, Port & International Logistics*" funded by the Ministry of Land, Transport and Maritime Affairs, Government of South Korea.

References

1. Fox MS, Barbyceanu M, Teigen R (2000) Agent-oriented supply chain management. Int J Flex Manuf Sys 1:165–188
2. Dyson RG (2000) Performance measurement and data envelopment analysis—ranking are ranks! OR Insight 13(1):3–8
3. Seiford L (2001) Data envelopment analysis: the evolution of the state of the art (1978–1995). J Productiv Anal 1(1):1–17
4. Liang L, Yang F, Cook WD, Zhu J (2006) DAE models for supply chain efficiency evaluation. Ann Oper Res 145(1):35–49
5. Reiner G, Hofmann P (2006) Efficiency analysis of supply chain processes. Int J Prod Res 144(23):5065–5087
6. SCOR, Supply Chain Council [Online]; http://www.supply-chain.org
7. Cheung KL, Hansman WH (2004) An exact performance evaluation for the supplier in a two-echelon inventory system. Oper Res 48(1):646–658
8. Chen Y, Zhu J (2001) Measuring information technology's indirect impact on firm performance. Inf Tech Manag J 17(1):1–17

9. Golany B, Roll Y (1989) An application procedure for DEA. OMEGA 17(3):237–250
10. Zhu J (2003) Quantitative models for performance evaluation and benchmarking: data envelopment analysis with spreadsheets. Kluwer Academic Publishers, Boston
11. Fare R, Grosskopf S (2000) Network DEA. Soc Econ Plann Sci 34:35–49
12. Castelli L, Pesenti R, Ukovich W (2004) DEA-like models for the efficiency evaluation of hierarchically structured units. Eur J Oper Res 154:465–476
13. Chopra S, Meindl P (2001) Supply chain management: strategy planning operation. Prentice Hall, New Jersey
14. Sharma MJ, Yu SJ (2010) Capturing the process efficiency and congestion of supply chains. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, pp 2429–2442
15. Sharma MJ, Yu SJ (2010) Benchmark optimization and attribute identification for improvement of container terminals. Eur J Oper Res 201(2):568–580
16. The 21st century retail supply chain: the key imperatives for retailers (2009) Abeerden Report
17. Cooper WW, Seiford LM, Zhu J (2004) Hand book on data envelopment analysis. Kluwer Academic Publishers, Boston
18. Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making unit. Eur J Oper Res 2:429–444
19. Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale efficiencies in data envelopment analysis. Manag Sci 30(9):1078–1092
20. Kao C, Hwang SN (2008) Efficiency decomposition in two-stage data envelopment analysis: an application to non-life insurance companies in Taiwan. Eur J Oper Res 185(1):418–429
21. Chen Y, Cook WD, Li N, Zhu J (2009) Additive efficiency decomposition in two-stage DEA. Eur J Oper Res 196:1170–1176
22. Fare R, Grosskopf S, Lovell CAK (1994) Production frontiers. Cambridge, England
23. Fare R, Grosskopf S (1983) Measuring congestion in production. Z. Nationalökon 43:251–271
24. Davenport TH, Short JE (1990) The new industrial engineering: information technology and business process redesign. Sloan Manag Rev 31(4):11–27

Chapter 40
Comparison of Dry and Flood Turning in Terms of Dimensional Accuracy and Surface Finish of Turned Parts

Noor Hakim Rafai and Mohammad Nazrul Islam

Abstract This work presents experimental and analytical results of a comparison of dry and flood turning in terms of dimensional accuracy and surface finish of turned parts. Subsequently, the influence of independent input parameters on dimensional accuracy and surface finish characteristics is investigated in order to optimize their effects. Three techniques—traditional analysis, Pareto ANOVA, and the Taguchi method—are employed. Hardened alloy steel AISI 4340 has been selected as work material. The results show that for certain combinations of cutting parameters, dry turning produced better dimensional accuracy compared to that produced by flood turning. Therefore, in the future, it will be possible to develop a system through modelling the cooling process that will be capable of predicting the situations where dry turning will be beneficial. This will reduce the application frequency of cutting fluids by avoiding their unnecessary applications and, consequently, their negative impact on the environment.

40.1 Introduction

Turning, in which material is removed from the external surface of a rotating workpiece, is one of the most basic material removal processes. Therefore, turning is the first choice for machining cylindrical parts. The performance of a turning

N. H. Rafai (✉)
Department of Manufacturing and Industrial Engineering, Universiti Tun Hussein Onn Malaysia (UTHM), 86400, Parit Raja, Batu Pahat, Johor, Malaysia
e-mail: nhakim@uthm.edu.my

M. N. Islam
Department of Mechanical Engineering, Curtin University, GPO Box U1987, Perth, WA 6845, Australia
e-mail: m.n.islam@curtin.edu.au

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_40,
© Springer Science+Business Media B.V. 2011

operation is greatly influenced by the application of cutting fluid, and, in this regard, turning operations can be classified into different types, such as dry turning, turning with minimum quantity lubrication (MQL), flood turning, and cryogenic turning. Of these, *flood turning* is the most traditional technique and by far the most widely used in industry. The process is characterized by the application of a large quantity of liquid, known as *cutting fluid*, at the cutting tool and workpiece interface.

In flood turning, also known as *wet turning*, cutting fluid is applied for a number of reasons, such as to reduce the cutting temperature, to lengthen the tool life, to produce a better surface finish, to improve dimensional accuracy, and to facilitate chip disposal. However, in recent years, the application of cutting fluids in machining operations has attracted immense scrutiny due to its adverse effects on health and the environment. Consequently, dry turning has gained renewed interest for its potential environmental and economic benefits. Nevertheless, in spite of all its economic and environmental benefits, the dimensional accuracy and surface finish of component parts produced by dry turning should not be sacrificed. Therefore, to make an informed decision, a direct comparison is needed between dry and flood turning in terms of the dimensional accuracy and surface finish of turned parts.

Comparison between various forms of turning has received attention in the research literature. A number of studies have reported on comparisons of different types of turning operations in terms of various machining performance characteristics such as tool wear [1–3], cutting force [4], cutting temperature [4], and productivity [5]. Comparatively, there have been relatively few studies reporting on the dimensional accuracy of turned parts [6], even though, in the majority of cases, dimensional accuracy is the first criterion in determining the acceptability of manufactured parts. Nonetheless, a number of publications [3, 6] have reported on comparisons of surface roughness achievable by different types of turning operations. In [7], an attempt was made to compare the shortcomings of dry turning over flood turning by monitoring the dimensional accuracy and surface finish characteristics of turned parts. The work presented here is an extended and revised version of our previous work [7].

Turning, like any other machining process, is greatly influenced by independent input variables—cutting speed, feed rate, and depth of cut—commonly known as *cutting conditions*. These cutting conditions are also believed to have significant effects on the dimensional accuracy and surface finish of the machined parts. Since these cutting parameters can be chosen and controlled by the user and are reflected in the dimensional accuracy and surface finish of machined parts, they are the appropriate parameters for investigating their effects on the quality of turned parts. Therefore, the objective of this research is to compare dry and flood turning in terms of the dimensional accuracy and surface finish of turned parts and to investigate the influences of independent input parameters (cutting conditions) on quality characteristics in order to optimize their effects.

40.2 Dimensional Accuracy and Surface Finish of Turned Parts

The dimensional accuracy of turned parts are specified by a number of dimensional accuracy characteristics; of these, diameter error and circularity are the two most important. Thus, in this study, they are selected for monitoring the dimensional accuracy of turned parts. *Diameter error* is the difference between the measured diameter and the designed diameter, where a positive error indicates undercutting of a cylindrical workpiece. It is an important dimensional characteristic of turned component parts, especially when cylindrical fits are involved.

For turned parts, circularity (also known as roundness or out-of-roundness) is another important dimensional accuracy characteristic that is geometric in nature. It is particularly important for rotating component parts where excessive circularity values may cause unacceptable vibration and heat. *Circularity* is defined by two concentric circular boundaries within which each circular element of the surface must lie [8].

Surface roughness is another important quality characteristic that can dominate the functional requirements of many component parts. For example, a good surface roughness value is necessary to prevent premature fatigue failure; to improve corrosion resistance; to reduce friction, wear, and noise; and, finally, to improve product life. *Surface roughness* is a measure of the fine irregularities on a surface and is defined by the height, width, direction, and shape of irregularities. There are a number of parameters currently available for measuring the surface roughness value. Yet, no single parameter appears to be capable of describing the surface quality adequately. In this study, *arithmetic average*, a height parameter, has been adopted to represent surface roughness, since it is the most frequently used and internationally accepted parameter.

40.3 Scope

The main objectives of this project are to investigate the deficiency of dry turning compared to flood turning in terms of the dimensional accuracy and surface finish of turned parts and to explore how these quality characteristics are influenced by the three independent input variables: cutting speed (A), feed rate (B), and depth of cut (C). To achieve this goal, a three-level three-parameter experiment was designed using design-of-experiment methodology. Two sets of experiments, each with 27 runs, were conducted under both dry and flood conditions. The dimensional accuracy and surface finish characteristics of the resulting turned parts were then checked using a general purpose coordinate measuring machine (CMM) and a surface finish analyzer. The results are analyzed by three techniques: (1) traditional analysis, (2) Pareto analysis of variation (ANOVA), and (3) Taguchi's signal-to-noise ratio (S/N) analysis.

In the traditional analysis, the mean values of the measured variables are used. This tool is particularly suitable for monitoring a trend of change in the relationship of variables.

Pareto ANOVA is an excellent tool for determining the contribution of each input parameter and their interactions with the output parameters (dimensional accuracy and surface finish characteristics). It is a simplified ANOVA method that does not require an ANOVA table; further details of Pareto ANOVA can be found in [9].

For the Taguchi method, the *signal-to-noise ratio* was calculated using the following formula [10]:

$$S/N = -10 \log \frac{1}{n} \left(\sum_{i=1}^{n} \frac{1}{y_i^2} \right) \quad (1)$$

where *S/N* is the signal-to-noise ratio (in *dB*), n is the number of observations, and y is the observed data.

The above formula is suitable for quality characteristics in which the adage "the smaller the better" holds true. This is the case for all three quality characteristics considered. The higher the value of the *S/N* ratio, the better the result is because it guarantees optimum dimensional accuracy and surface finish with minimum variance. A thorough treatment of the Taguchi method can be found in [10].

40.4 Experimental Work

The experiments were planned using Taguchi's orthogonal array methodology [10]. A three-level L_{27} orthogonal array was selected for our experiments. For each type of turning, 9 parts were produced. Each part was divided into three segments. Each segment was turned with the cutting conditions determined by design of experiment (DoE). The position of the segments was allocated randomly. The total number of experiments was 54. The values of three input cutting parameters (cutting speed, feed rate, and depth of cut) were selected on the basis of the capacity and limitation of the lathe machine used; details are given in Table 40.1.

Hardened round bars of AISI 4340 with 30 HRC hardness value, 40 mm nominal diameter, and 120 mm length were used as blank material. AISI 4340 is a hard and difficult-to-machine material; it was selected anticipating larger differentiations in dimensional accuracy and surface finish characteristics when

Table 40.1 Input variables

Input parameters	Unit	Symbol	Levels		
			Level 0	Level 1	Level 2
Cutting speed	m/min	*A*	54	150	212
Feed rate	mm/rev	*B*	0.11	0.22	0.33
Depth of cut	mm	*C*	0.50	1.00	1.50

machining difficult-to-machine work materials. Parts were produced on a three-axis CNC turning centre, GT-250MA, manufactured by Yeong Chin Machinery Industries Co. Ltd. (YCM), Taiwan. For holding the workpiece, a three-jaw chuck supported at the dead center was employed. Square-shaped inserts with enriched cobalt coating manufactured by Stellram, USA, were used as the cutting tools. The inserts were mounted on a standard PSDNN M12 tool holder. Castrol Clearedge EP690, a semi-synthetic soluble cutting fluid, was used in flood turning.

The precision measurements were taken by a Discovery Model D-8 coordinate measuring machine (CMM), manufactured by Sheffield, UK. The probes used were spherical probes with a star configuration, manufactured by Renishaw Electrical Ltd. The diameters of test parts were calculated using the standard build-in software package of the CMM. Eight points were measured for each measurement of diameter, and each measurement was repeated three times. The circularity data was also obtained from the CMM. The surface roughness parameter arithmetic average (R_a) for each surface was determined by a surface-measuring instrument: the Surftest SJ-201P, manufactured by Mitutoyo, Japan.

40.5 Results and Analysis

All the parts were checked thoroughly, and an enormous amount of data was collected and subsequently analyzed. Although in the analysis of the work, all these relationships were considered at different stages, due to space constraints, only a few are illustrated.

40.5.1 Diameter Error

A comparison of diameter error for dry and flood turning for different cutting conditions is illustrated in Fig. 40.1. The figure shows that at a low cutting speed, there is no noticeable difference between dry and flood turning. At a medium cutting speed and low feed rate (A_1B_0), flood turning produces the most benefit. As the feed rate is increased, the diameter error for dry turning is improved, whereas for flood turning, diameter error is deteriorated, and at a medium cutting speed and high feed rate (A_1B_2) both graphs converge. When both cutting speed and feed rate are further increased, the two graphs remain nearly identical. Nevertheless, at a high cutting speed and high feed rate (A_2B_2), dry turning results in better quality in terms of diameter error.

The Pareto ANOVA for diameter error for dry and flood turning is given in Fig. 40.2. The figure shows that in both cases, cutting speed (A) has the most significant effect on diameter error, and the contribution ratios for both cases are about the same ($P \cong 32\%$). The interaction between cutting speed and feed rate ($A \times B$) plays a significant role in both cases, although the influence of the interaction between cutting speed and feed rate is higher in flood turning ($P = 28.3\%$)

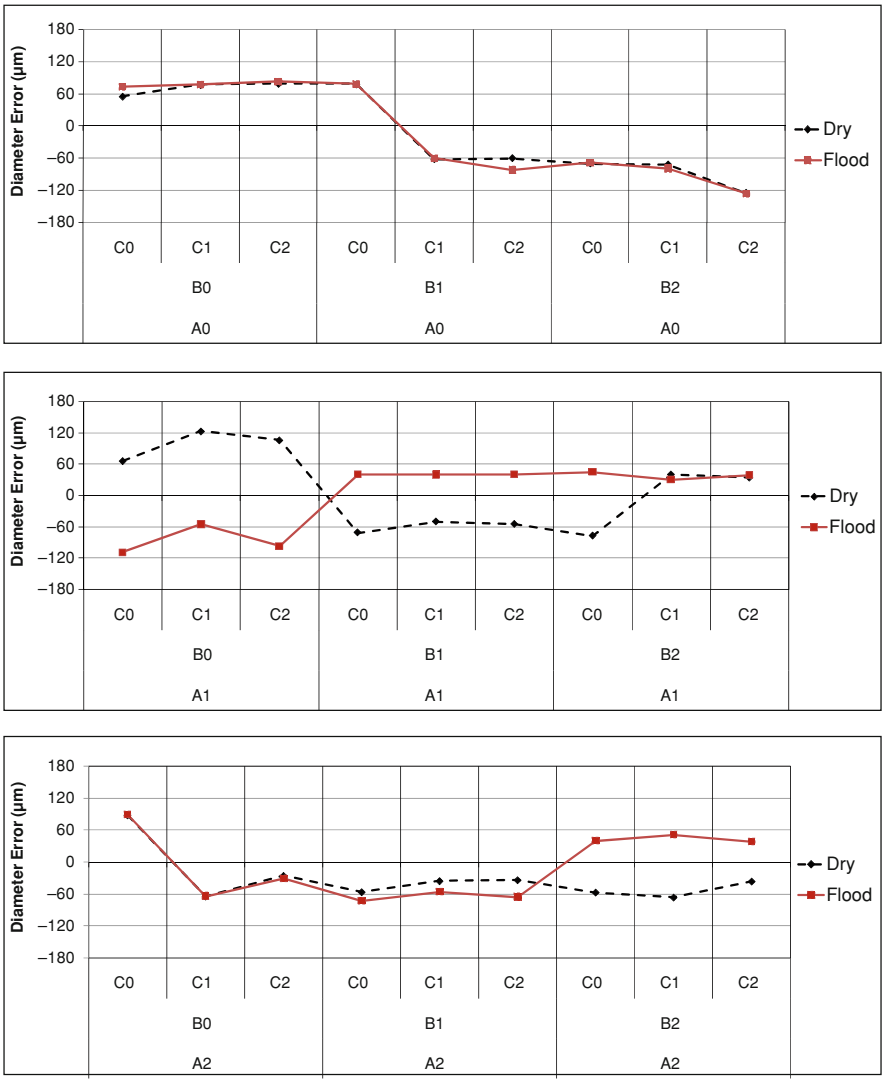


Fig. 40.1 Comparison of diameter error for dry and flood turning under different cutting conditions

than dry turning ($P = 19.7\%$). It is interesting to note that the contribution of feed rate (B) and depth of cut (C) approximately swaps with each other. In dry tuning, the contribution ratio for feed rate and depth of cut are $P = 2.8$ and 14.7%, respectively, whereas in flood turning, the contribution ratio for feed rate and depth of cut are $P = 13.6$ and 3.6%, respectively.

The response graphs for diameter errors for dry and flood turning are illustrated in Fig. 40.3. As the slopes of the response graphs represent the strength of contribution, the response graphs confirm the findings of Pareto AVOVA analysis.

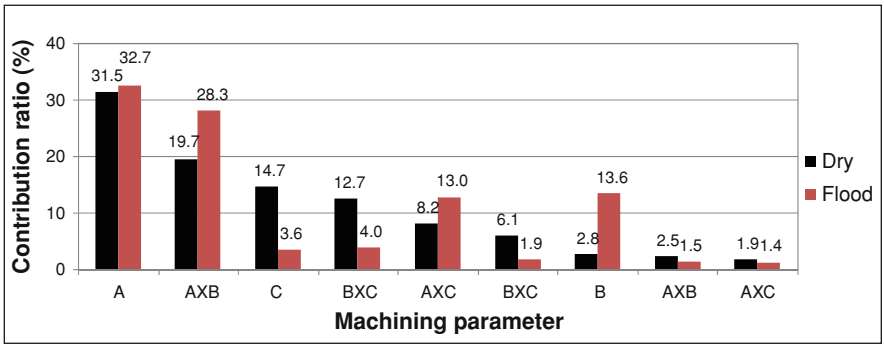


Fig. 40.2 Comparison of Pareto ANOVA analyses for diameter error for dry and flood turning

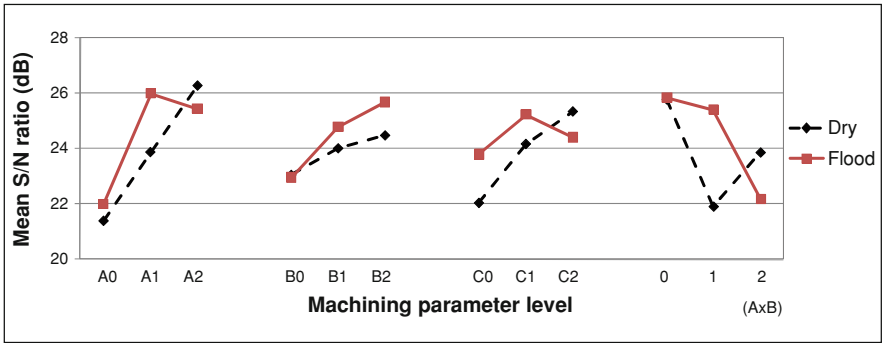


Fig. 40.3 Comparison of response graphs for diameter error for dry and flood turning

Figure 40.3 also reveals that with the increase of each main cutting parameter, the *S/N* ratio is increased, whereas this trend is changed in flood turning.

The experimental results presented above are difficult to explain because cutting speed, the most dominant factor (Fig. 40.2), can affect diameter error in a number of ways, such as by changing elastic deformation of a workpiece, induced by a change of cutting force, by tool wear, by an increase of thermal distortion, by the formation of built-up edge (BUE), and by an increase of radial spindle error. Furthermore, there is strong interaction between cutting speed and feed rate ($A \times B$), which makes definite conclusions difficult.

40.5.2 Circularity

A comparison of circularity for dry and flood turning for different cutting conditions is illustrated in Fig. 40.4. The figure illustrates that at a low cutting speed, dry turning produced better quality in terms of the circularity of turned parts. As the feed rate is increased, the circularity for flood turning is improved, whereas

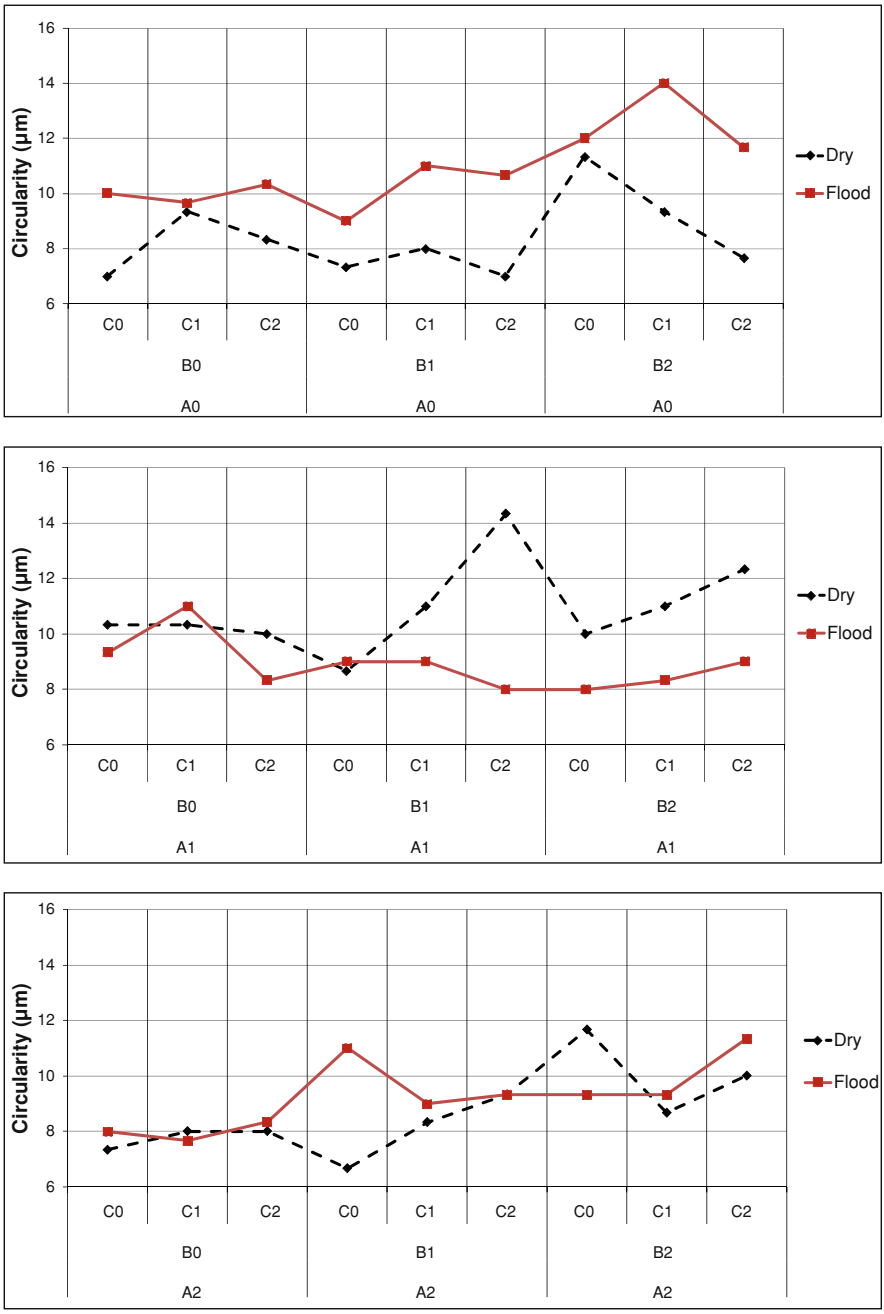


Fig. 40.4 Comparison of circularity for dry and flood turning under different cutting conditions

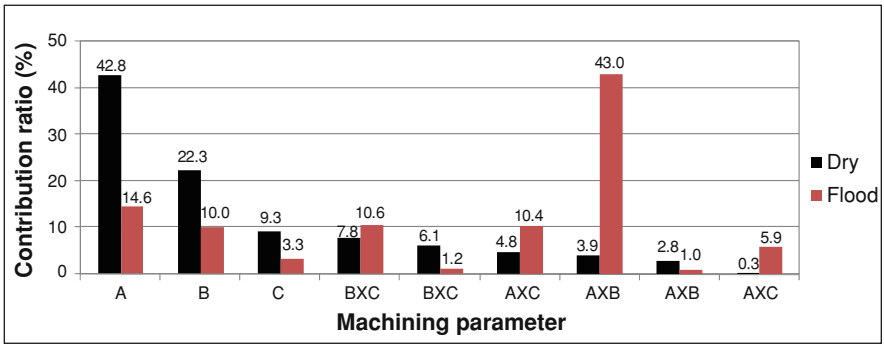


Fig. 40.5 Comparison of Pareto ANOVA analyses for circularity for dry and flood turning

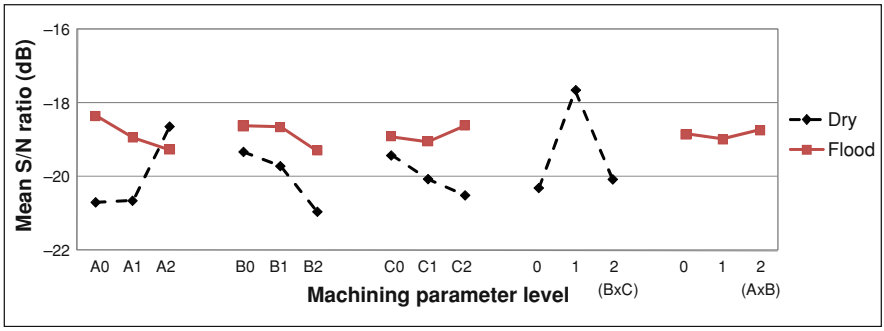


Fig. 40.6 Comparison of response graphs for diameter error for dry and flood turning

for dry turning, circularity is deteriorated. At a medium cutting speed and high feed rate (A_1B_2), flood turning produced the most benefit. At a high cutting speed, region results are inconclusive.

The Pareto ANOVA for circularity for dry and flood turning is given in Fig. 40.5. The figure shows that for dry turning, cutting speed has the highest contribution ratio ($P = 42.8\%$), whereas for flood turning, the interaction between cutting speed and feed rate ($A \times B$) have the most significant effect ($P = 43.0\%$). It is worth pointing out that the total of all interaction effects is about two times larger for flood turning ($P \cong 72\%$) than for dry turning ($P \cong 36\%$). This indicates that optimizing the circularity of turned parts in flood turning by three independent cutting parameters will be difficult.

The response graphs for circularity for dry and flood turning are illustrated in Fig. 40.6. The figure shows that the application of cutting fluid changes the direction response graphs for the main interaction effects—cutting speed (A) and depth of cut (C). The main interaction effect parameter also changes from ($B \times C$) in dry turning to ($A \times B$) in flood turning.

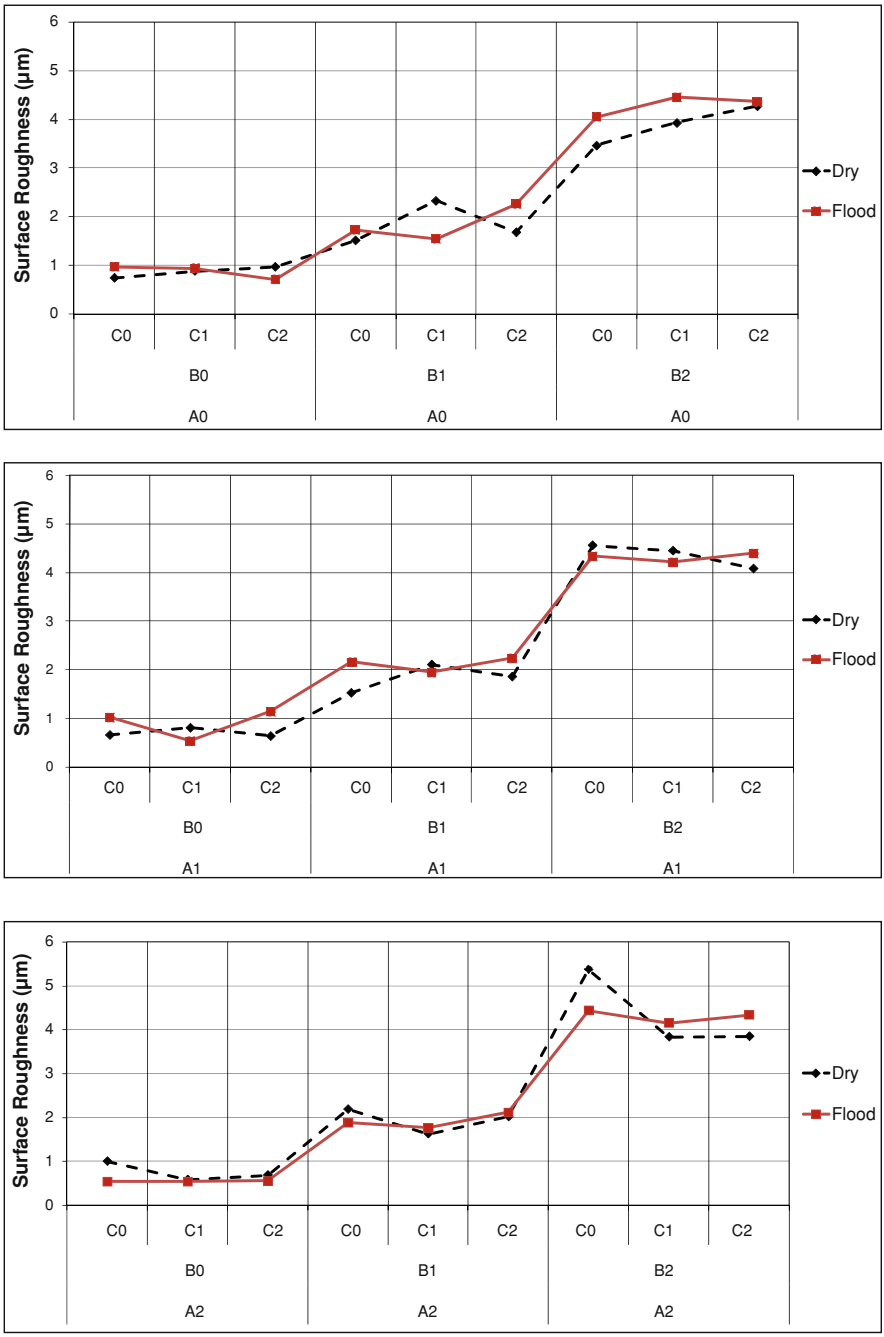


Fig. 40.7 Comparison of surface roughness for dry and flood turning under different cutting conditions

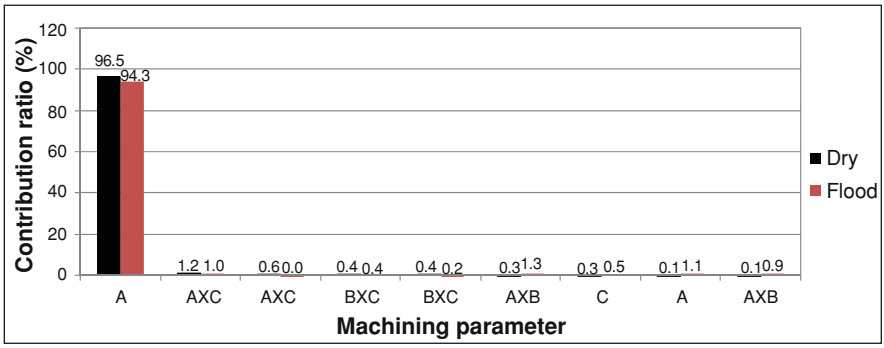


Fig. 40.8 Comparison of Pareto ANOVA analyses for surface roughness for dry and flood turning

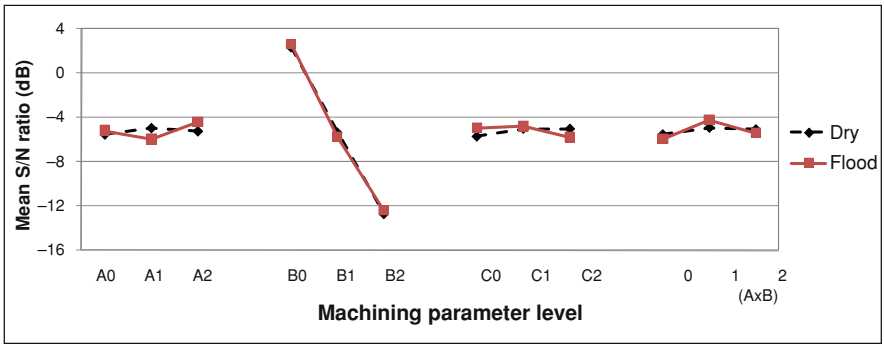


Fig. 40.9 Comparison of response graphs for surface roughness for dry and flood turning

40.5.3 Surface Roughness

A comparison of surface roughness resulting from dry and flood turning for different cutting conditions is shown in Fig. 40.7. As expected, with an increased feed rate, the surface roughness values are increased for both cases and all cutting conditions investigated. However, contrary to widely held assumptions, Fig. 40.7 illustrates no considerable benefit to flood turning compared to dry turning in terms of surface roughness; on the contrary, there is a trend of slightly better surface roughness produced by dry turning in a low cutting speed and high feed rate (A_0B_2) region.

The contribution ratios achieved though Pareto ANOVA for surface roughness for dry and flood turning is given in Fig. 40.8. The figure shows that, in both cases, feed rate (B) has the most significant effect on surface roughness. The influence of feed rate on surface roughness is well known; however, in both cases, the extent of this influence is surprisingly very high ($P = 95\%$ for both cases). This means that for a practical machining operation, the other two cutting variables—cutting speed

and depth of cut—can be overlooked, although the cutting speed selected must be high enough to avoid BUE. Being influenced by a single parameter, it is relatively uncomplicated to optimize surface roughness. This conclusion is confirmed by the response graphs illustrated in Fig. 40.9. Nevertheless, further analysis reveals that the utilization of a low feed rate can improve the surface roughness of turned component parts. However, a reduction of feed rate decreases the production rate and should be employed as a last resort. For example, surface roughness can be improved by increasing the tool nose radius, a parameter not included in this study.

40.6 Concluding Remarks

The results presented in this work show that for certain combinations of cutting parameters, dry turning produced better dimensional accuracy compared to that produced by flood turning. This indicates that, in the future, it will be possible, through modelling the cooling process, to develop a system for finding in which situations dry turning will be beneficial, thus reducing the application frequency of cutting fluids and, consequently, their negative impact on the environment. The results also show that no considerable difference in surface roughness is produced by dry and flood turning. Some clear trends that appear in the traditional analyses are difficult to explain. Therefore, further research is needed to investigate these trends.

The results presented in this work should be treated with caution because other than the selected cutting parameters, there are many other factors, such as work materials, tool materials, tool geometry, and machine condition that may influence the outcome. These factors we intend to study in our future work.

References

1. Marksberry PW, Jawahir IS (2008) A comprehensive tool-wear/tool-life model in the evaluation of NDM (near dry machining) for sustainable manufacturing. Int J Mach Tools Manuf 48:878–886
2. Dhar NR, Kamruzzaman M, Ahmed M (2006) Effect of minimum quantity lubrication (MQL) on too wear and surface roughness in turning AISI-4340 steel. J Mater Process Technol 172:299–304
3. Sarma DK, Dixit US (2007) A comparison of dry and air-cooled turning of grey cast iron with mixed oxide ceramic tool. J Mater Process Technol 190:160–172
4. Varadaarajan AS, Philip PK, Ramamoorthy B (2002) Investigations on hard turning with minimal cutting fluid application (HTMF) and its comparison with dry and wet turning. Int J Mach Tools Manuf 42:193–200
5. Harrma VS, Dogra M, Suri NM (2009) Cooling techniques for improved productivity in turning. Int J Mach Tools Manuf 49:435–453
6. Dhar NR, Islam M, Islam S, Mithun MAH (2006) The influence of minimum quantity lubrication (MQL) on cutting temperature, chip and dimensional accuracy in turning AISI-1040 steel. J Mater Process Technol 171:93–99

7. Rafai NH, Islam MN (2010) Comparison of dry and flood turning in terms of quality of turned parts, In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July, 2010, London, UK, pp 2044–2049
8. ASME Y14.5-2009 (2009) Dimensioning and tolerancing, ASME, New York
9. Ross PJ (1988) Taguchi techniques for quality engineering. McGraw-Hill, New York
10. Park SH (1996) Robust design and analysis for quality engineering. Chapman & Hall, London

Chapter 41
Coordinated Control Methods of Waste Water Treatment Process

Magdi S. Mahmoud

Abstract This paper develops coordinated control methods for the regulation policies of high-strength waste water treatment process. A dynamic model for the activated sludge process with waste waters is presented and linearized for control studies. The control strategy regulates the feed rate to maintain a constant optimal substrate concentration in the reactor, which in turn minimizes the reaction time. The coordinated control method consists of three components: two components have independent direct effects on the behavior of the aerated basin and the settling tank and the third component coordinates the overall operation. Simulation results show the effectiveness of the developed methods.

41.1 Introduction

The use of the activated sludge process (ASP) for biological nitrogen removal has increased in many countries. This is a result of stricter effluent demands. To remove the nitrogen from the waste water in an ASP, the biological processes, nitrification and de-nitrification, are needed. In aerobic compartments, ammonium may be converted into nitrate (nitrification) and in anoxic compartments nitrate may be converted into gaseous nitrogen (de-nitrification). For the operation to work well, a sufficiently high concentration of dissolved oxygen (DO) is needed

M. S. Mahmoud (✉)
Systems Engineering Department, King Fahd University of Petroleum and Minerals, P.O. Box 5067, Dhahran 31261, Saudi Arabia
e-mail: msmahmoud@kfupm.edu.sa

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_41,
© Springer Science+Business Media B.V. 2011

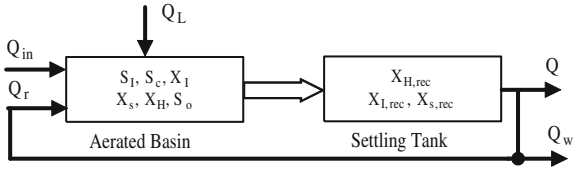


Fig. 41.1 Waste water treatment process

together with a sufficiently large aeration volume [1, 2]. A comprehensive overview of different approaches to DO control is given in [3].

It is known that activated sludge systems are affected by several dynamic variables which have influence over the output concentration of operation parameters. Fluctuation of temperature, flow and organics modifies the performance of the process and makes steady-state models inefficient for explaining the normal perturbations in waste water treatment plants. Under this situation, dynamic models are clearly in advantage.

The progressive deterioration of the water resources and the great quantity of polluted water produced in the industrial companies, give to the waste water treatment a great importance in the safeguarding of water's quality. So the monitoring of this kind of process has become an important task. The heart of this process is composed of two basins: the aerated basin and the settling tank (Fig. 41.1). The work reported in this chapter is a revised version of [4].

41.2 Mathematical Model

The fundamental phase of mathematical modeling consists in determining the reaction rates of the macroscopic variables of the system. The objective of the activated sludge process is to achieve, at a minimum cost, a sufficiently low concentration of biodegradable matter in the effluent, together with minimal sludge production. There is no reaction in the settling tank which delivers purified water after the decantation of sludge. A part of this later is recycled in the aerated basin. The fundamental phase of the mathematical modeling consists in determining the reaction rates of the macroscopic variables of the system to know the rate of biomass growth, substrate degradation and dissolved oxygen uptake. The second stage makes it possible to determine the system equations whose states variables are the concentrations in micro-organisms, substrate, recycled biomass and dissolved oxygen. These variables as well as the inputs and the outputs are gathered in mathematical expressions thus constituting the process model [5]. The mathematical model for the activated sludge process (aerated basin and settling tank) is based on the equations, resulting from mass balance considerations, carried out on each of the reactant of the process. The initial system is composed of nine states, four inputs and six outputs.

41.2.1 Aerated Basin Model

$$\begin{aligned}
\frac{dS_1}{dt} &= \frac{Q_{in}}{V_r}(S_{1,in} - S_1),\\
\frac{dS_s}{dt} &= \frac{Q_{in}}{V_r}(S_{3,in} - S_3) - \frac{1}{Y_H}\rho_1 + \rho_3,\\
\frac{dX_1}{dt} &= \frac{Q_{in}}{V_r}(X_{1,in} - X_1) + \frac{Q_r}{V_r}(X_{3,rec} - X_H) + (1 - f_{X_1})\rho_2 - \rho_3,\\
\frac{dX_s}{dt} &= \frac{Q_{in}}{V_r}(X_{3,in} - X_3) + \frac{Q_r}{V_r}(X_{3,rec} - X_H) + (1 - f_{X_1})\rho_2 - \rho_3,\\
\frac{dX_H}{dt} &= \frac{Q_{in}}{V_r}(X_{H,in} - X_H) + \frac{Q_r}{V_r}(X_{H,rec} - X_H) + \rho_1 - \rho_2,\\
\frac{dS_0}{dt} &= \frac{Q_{in}}{V_r}(S_{0,in} - S_0) + Q_L(C_3 - C_0) + \frac{1 - Y_H}{Y_H}\rho_1.
\end{aligned} \tag{41.1}$$

41.2.2 Settling Tank

$$\begin{aligned}
\frac{dX_{H,rec}}{dt} &= \frac{Q_{in} + Q_r}{V_{dec}}X_H - \frac{Q_r + Q_w}{V_{dec}}X_{H,rec},\\
\frac{dX_{I,rec}}{dt} &= \frac{Q_{in} + Q_r}{V_{dec}}X_I - \frac{Q_r + Q_w}{V_{dec}}X_{I,rec},\\
\frac{dX_{s,rec}}{dt} &= \frac{Q_{in} + Q_r}{V_{dec}}X_s - \frac{Q_r + Q_w}{V_{dec}}X_{s,rec},
\end{aligned} \tag{41.2}$$

where the parameters ρ_1, ρ_2 and ρ_3 are given by $\rho_1 = \mu_{max}\frac{S_s}{(K_s+S_s)(K_0+S_0)}$, $\rho_2 = b_H X_H$, $\rho_3 = K_h \frac{X_s X_H}{(K_s X_H + X_s)}\frac{S_0}{(K_0+S_0)}$.

From a control engineering standpoint, handling model (41.1)–(41.2) is quite hard [5] and therefore we direct attention to an alternative approaches. In terms of $x \in R^9$, $u \in R^4$ and $y \in R^6$ we cast the model equations (41.1)–(41.2) into the format

$$\frac{dx}{dt} = \dot{x} = f(x, u), \quad y = Cx. \tag{41.3}$$

Letting the system equilibrium point x_e, u_e be defined by $\frac{dx}{dt}|_{x=e,u=ue}$. Performing a linearization of system (41.3) around the equilibrium point yields the linearized model

$$\frac{dz}{dt} = \dot{z} = Az + Bv, \quad w = Cz, \tag{41.4}$$

$$A = \frac{\partial f}{\partial x}, \quad B = \frac{\partial f}{\partial u}, \qquad z = x - x_e, \quad v = u - u_e, \quad w = y - y_e \tag{41.5}$$

into consideration the two subsystems (aerated basin and settling tank), we express model (41.4)–(41.5) in the form

$$\frac{d}{dt}z_1 = A_1 z_1 + B_1 v_1 + g_1, \tag{41.6}$$

$$g_1 = A_2 z_2, \tag{41.7}$$

$$\frac{d}{dt}z_2 = A_2 z_2 + B_2 v_2 + g_2, \tag{41.8}$$

$$g_2 = A_3 z_1, \tag{41.9}$$

where

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}, \quad \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}, \quad \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix},$$
$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

The numerical values of the respective matrices are given by

$$A_1 = \begin{bmatrix} 30 & 0 & 0 & 0 & 0 & 0 \\ 0 & -14.5 & 0 & 166 & -150 & 0 \\ 0 & 0 & -29 & 0 & 1060 & 0 \\ 0 & 0 & 0 & -29 & 1120 & 0 \\ 0 & -14.5 & 0 & 0 & -29 & 0 \\ 0 & -14.5 & 0 & 0 & 0 & -12.5 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 68.03 & 0 \\ 0 & 0 & 68.03 \\ 14.5 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & -1136 & 0 \\ 0 & 0 & -1136 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1136 & 0 & 0 \end{bmatrix},$$

$$A_4 = \begin{bmatrix} 83.77 & 0 & 0 \\ 0 & 83.77 & 0 \\ 0 & 0 & 83.77 \end{bmatrix}, \quad B_1 = \begin{bmatrix} -5 & 0 \\ 5 & 0 \\ 0 & -540 \\ 8 & -314 \\ -13 & 0 \\ 0 & -620 \end{bmatrix},$$

$$B_2 = \begin{bmatrix} -98 & -583 \\ 0 & -339 \\ 0 & -670 \end{bmatrix}, \quad C_1 = I_6, \quad C_2 = I_3.$$

We recall that the linearized model (41.6)–(41.9) represents an interconnected system with linear coupling pattern [6]. In the next section, we present two distinct decentralized control methods to regulate the dynamic behavior of the waste water treatment process.

41.3 Coordinated Control Methods

Employing the linear quadratic control theory and considering the decoupled case $g_1 = 0,\ g_2 = 0$, we optimize each subsystem

$$\frac{d}{dt} z_1 = A_1 z_1 + B_1 v_1, \tag{41.10}$$

$$\frac{d}{dt} z_2 = A_2 z_2 + B_2 v_2 \tag{41.11}$$

with respect to the quadratic performance indices

$$J_s = \int_0^\infty e^{2\pi t} [z_s^t Q_s z_s + v_s^t R_s v_s]\, dt, \quad s = 1, 2, \tag{41.12}$$

where Q_s is an rXr symmetric, nonnegative definite matrix ($r = 6$ for subsystem 1 and $r = 3$ for subsystem 2), R_s is an tXt is an symmetric, positive definite matrix ($t = 2$ for subsystems 1 and 2), and π is a nonnegative number. As known [7], under the assumption that the pair (A_s, B_s) is completely controllable, there exists a unique optimal control law

$$v_s^* = -K_s^* z_s = -R_s^{-1} B_s^t P_s z_s, \quad s = 1, 2 \tag{41.13}$$

and P_s is an rXr symmetric, positive definite matrix which is the solution of the Riccati equation

$$P_s(A_s + \pi_s I) + (A_s + \pi_s I)^t P_s - P_s B_s R_s^{-1} B_s^t P_s + Q_s = 0, \quad s = 1, 2 \tag{41.14}$$

such that $v*_s$ minimizes J_s in (41.12). The associated optimal cost is

$$J_s = z_s^t(0) P_s z_s(0), \quad s = 1, 2. \tag{41.15}$$

Under the assumption that Q_s can be factored as $C_s C_s^t$, where C_s is an pXp constant matrix, so that the pair (A_s, C_s) is completely observable, each closed-loop subsystem

$$\frac{d}{dt} z_s = (A_s - B_s R_s^{-1} B_s^t P_s) z_s, \quad s = 1, 2 \tag{41.16}$$

is globally exponentially stable with the degree π [7]. That is, the solution z_s of (41.16) approaches the equilibrium at the origin at least as fast as $e^{-\pi t}$ for all initial conditions.

41.3.1 Method 1

To compensate for the interactions $(g_1 \neq 0,\ g_2 \neq 0)$, and preserve the subsystems' autonomy, we apply the controls

$$v_s = v_s^* + v_s^c, \quad s = 1, 2. \tag{41.17}$$

The controls v_s^c are selected in the form $v_s^c = -K_s^c z_s,\ s = 1, 2$ where the gains K_s^c are compensatory controls to adjust the dynamic behavior of the aerated basin and settling tank subsystems. It follows from [6] that

$$v_s^c = \begin{bmatrix} v_1^c \\ v_2^c \end{bmatrix} = - \begin{bmatrix} R_1^{-1} B_1^t & 0 \\ 0 & R_2^{-1} B_2^t \end{bmatrix} \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \tag{41.18}$$

where

$$\begin{aligned}
F_{11} &= -[G_{11}^{-1} G_{12} G_{22}][K_2^* A_{21} - A_{12}^t K_1^*], \\
F_{21} &= -D_{22}[K_{12}^* A_{21} - A_{12}^t K_1^*], \\
F_{12} &= [G_{11}^{-1} + G_{11}^{-1} G_{12} D_{22} G_{21} G_{11}^{-1}][K_{12}^* A_{21} - A_{12}^t K_1^*], \\
G_{12} &= A_{21}^t, G_{21} = A_{12}^t, \\
F_{22} &= -[D_{22} G_{21} G_{11}^{-1}][K_1^* A_{12} - A_{21}^t K_2^*], \\
G_{11} &= [K_1^* B_1 R_1^{-1} B_1^t - A_1^t], \\
G_{22} &= [K_2^* B_2 R_2^{-1} B_2^t - A_2^t], \\
D_{22} &= [G_{22} - G_{21} G_{11}^{-1} G_{12}]^{-1}.
\end{aligned}$$

From (41.13), (41.18) and (41.19), the coordinating controls take the form

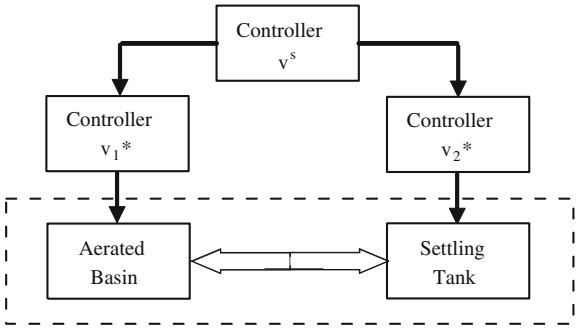


Fig. 41.2 Coordinated control structure

$$v_s = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = -R \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \tag{41.19}$$

and the coordinating gain matrix R is given by

$$R = -\begin{bmatrix} R_1^{-1} B_1^t (K_1^* + F_{11}) & R_1^{-1} B_1^t F_{12} \\ R_2^{-1} B_{12}^t F_{21} & R_2^{-1} B_2^t (K_2^* + F_{22}) \end{bmatrix} \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}. \tag{41.20}$$

This scheme is depicted in Fig. 41.2 where the controller scheme consists of two-levels: local controls at the lower level and coordinating control at the higher level.

41.3.2 Method 2

An alternative scheme is to constrain the gains K_s^c so as to neutralize the effect of coupling between the subsystems. It follows from [8] these gains can be computed by the formula

$$K_s^c = (BB^t)^{-1} B^t \begin{bmatrix} 0 & A_2 \\ A_3 & 0 \end{bmatrix}. \tag{41.21}$$

In the next section, we perform numerical simulation based on the coordinated schemes (41.18) and (41.19).

41.4 Simulations Results

Starting with the specification of the weighting matrices, we ran several computer simulations on the open loop nonlinear and linearized models. Satisfactory simulation results were attained using

$$Q = \text{diag}[10, 10, 10, 1, 0.1, 0.5, 1, 1, 0.01]$$

$$R = \text{diag}[1, 1, 1, 0.5].$$

Then we evaluate expressions (41.18) and (41.19). The difference between the two expressions was found to be small.

The gain matrices are given by

$$K_2^* = \begin{bmatrix} -28.7 & -0.31 & -1.01 & 1.65 & -0.85 & 0.32 \\ 4.45 & -0.12 & 0.57 & -1.75 & 0.67 & 0.07 \end{bmatrix},$$

$$K_2^* = \begin{bmatrix} 0.04 & 0.01 & -0.02 \\ -1.12 & 0.11 & -0.38 \end{bmatrix},$$

$$K^c = [\, K_1^c \quad K_2^c \,],$$

$$K_1^c = \begin{bmatrix} -8.6 & -0.15 & -1.01 & 0.56 & -0.33 \\ 1.55 & -0.04 & 0.09 & -0.27 & 0.06 \\ -6.43 & -0.05 & -0.01 & 0.02 & -0.01 \\ -7.35 & 0.13 & 0.05 & -1.08 & 0.27 \end{bmatrix},$$

$$K_2^c = \begin{bmatrix} 0.18 & 0.47 & 0.33 & -0.51 \\ 0.04 & 0.42 & -0.19 & -0.82 \\ 0.31 & 0.01 & 0.01 & 0.01 \\ -0.82 & 0.11 & 0.05 & 0.04 \end{bmatrix}.$$

The corresponding state and control trajectories are plotted in Figs. 41.3, 41.4, 41.5, 41.6, 41.7, 41.8, 41.9, 41.10, 41.11, 41.12, 41.13, 41.14, and 41.15. From the ensuing results, it is quite clear that the developed coordinated control methods have been effective in regulating the dynamic behavior of the waste water treatment variables.

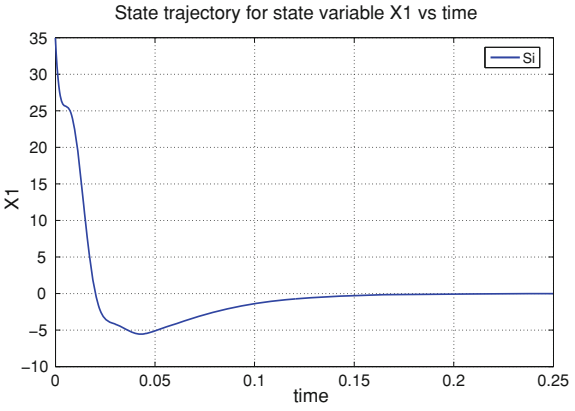


Fig. 41.3 Trajectory of soluble inert organic matter

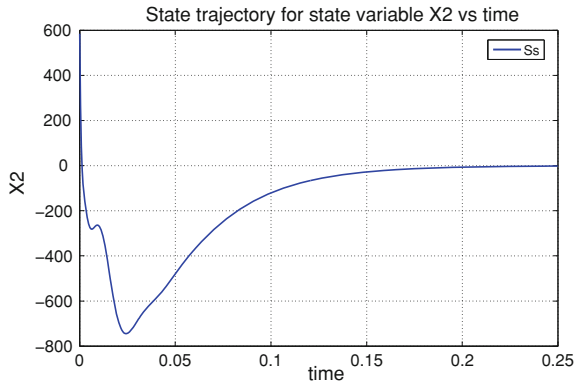


Fig. 41.4 Trajectory of readily biodegradable substrate

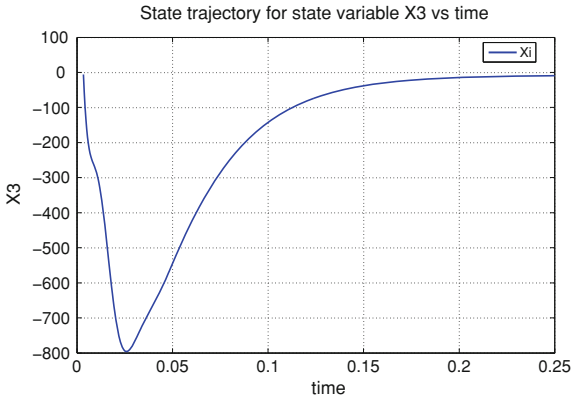


Fig. 41.5 Trajectory of particulate inert organic matter

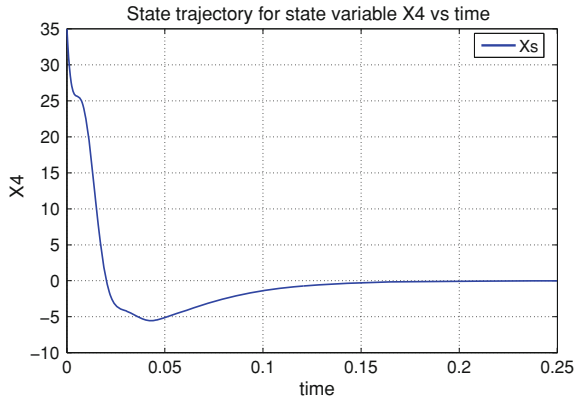


Fig. 41.6 Trajectory of slowly biodegradable substrate

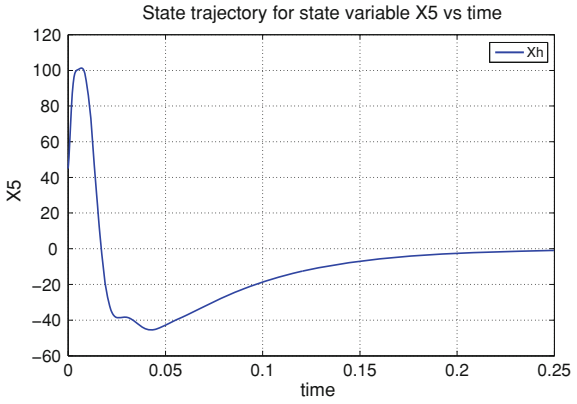


Fig. 41.7 Trajectory of heterotrophic biomass

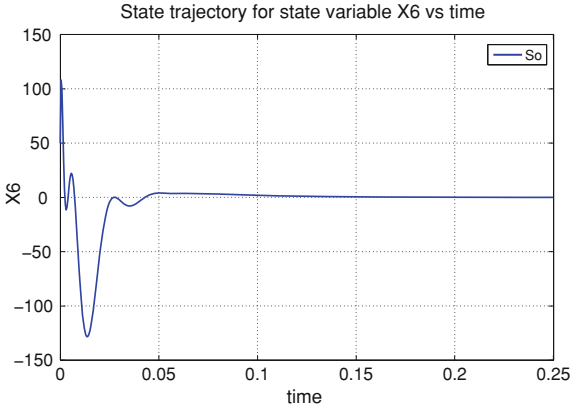


Fig. 41.8 Trajectory of oxygen dissolved in the diet

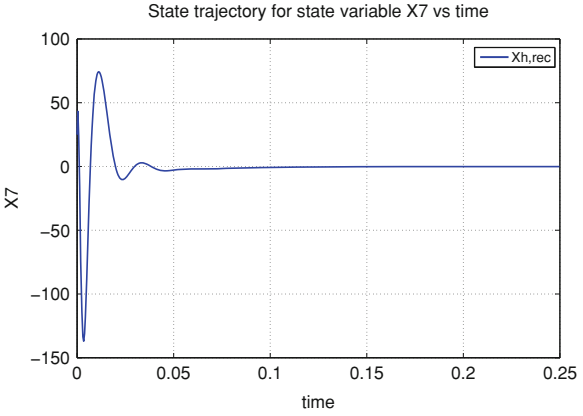


Fig. 41.9 Trajectory of heterotrophic biomass in the diet

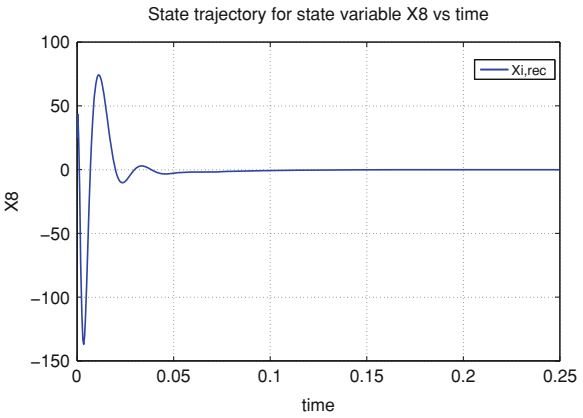


Fig. 41.10 Trajectory of organic matter recycled inert particulate

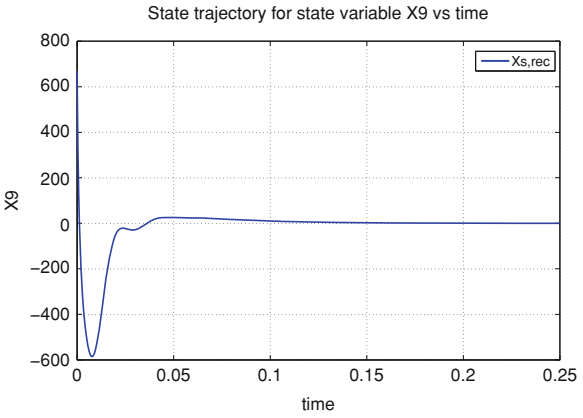


Fig. 41.11 Trajectory of slowly biodegradable substrate recycled

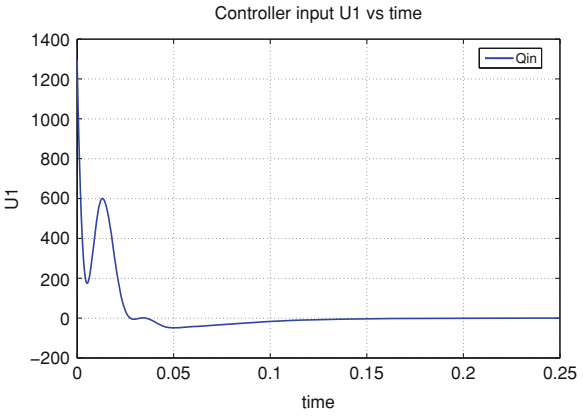


Fig. 41.12 Trajectory of inlet flow

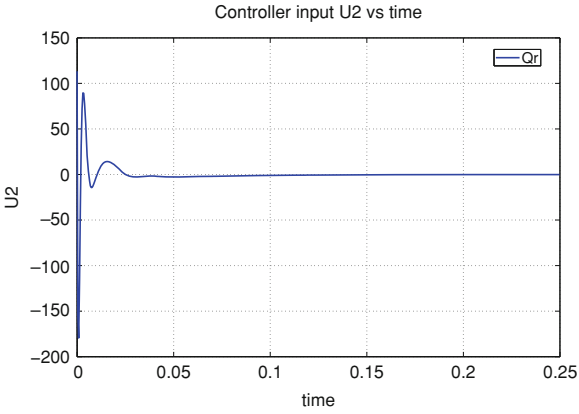


Fig. 41.13 Trajectory of flow recycling between the clarifier and the reactor

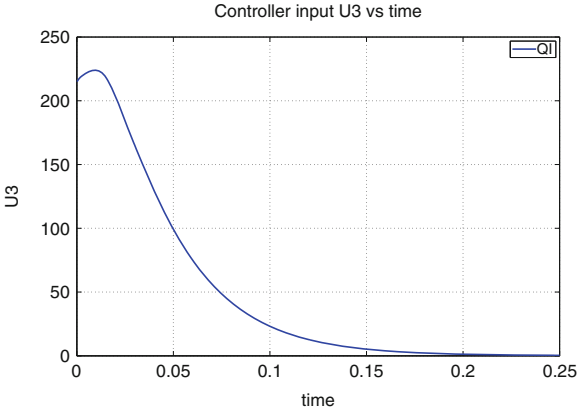


Fig. 41.14 Trajectory of air flow in the aeration tank

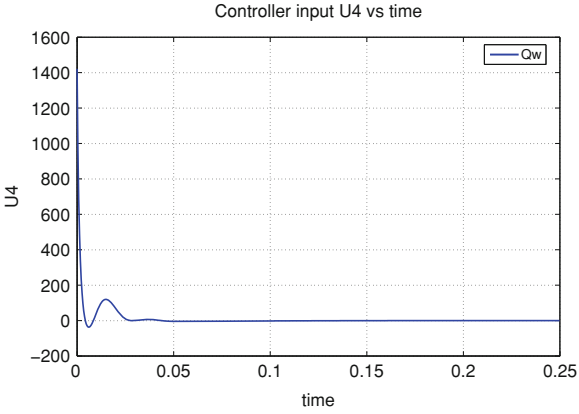


Fig. 41.15 Trajectory of flow purge

41.5 Conclusion

In this work, we have developed coordinated control methods for the regulation policies of high-strength waste water treatment process. A dynamic model for the activated sludge process with waste waters has been presented and linearized for control studies. The coordinated control method consists of three components: two components have independent direct effects on the behavior of the aerated basin and the settling tank and the third component coordinates the overall operation. Simulation results have shown the effectiveness of the developed methods.

Acknowledgments The work was supported by the KFUPM deanship of scientific research through project IN100018.

41.1 Appendix A: Variables and Model Parameters

S_i	Concentration of soluble inert organic matter (mg/l)
S_s	Concentration of readily biodegradable substrate (mg/l)
X_i	Concentration of particulate inert organic matter (mg/l)
X_s	Concentration of slowly biodegradable substrate (mg/l)
X_h	Concentration of heterotrophic biomass (mg/l)
$X_{h,rec}$	Concentration of recycled heterotrophic biomass (mg/l)
$S_{i,in}$	Concentration of soluble inert organic matter in the diet (mg/l)
$X_{I,rec}$	Concentration of organic matter recycled inert particulate (mg/l)
$X_{s,rec}$	Concentration of slowly biodegradable substrate recycled (mg/l)
$S_{i,in}$	Concentration of soluble inert organic matter in the diet (mg/l)
$S_{s,in}$	Concentration of readily biodegradable substrate in feed (mg/l)
$X_{i,in}$	Concentration of particulate inert organic matter in the diet (mg/l)
$X_{s,in}$	Concentration of slowly biodegradable substrate in feed (mg/l)
$X_{h,in}$	Concentration of heterotrophic biomass in the diet (mg/l)
$S_{0,in}$	Concentration of oxygen dissolved in the diet (mg/l)
Q_{in}	Inlet flow (l/h)
p_1	Speed specific heterotrophic growth (1/h)
p_2	Speed specific mortality of heterotrophic (1/h)
p_3	Speed specific hydrolysis of organic matter absorbed (1/h)
b_H	Coefficient of mortality of heterotrophic organisms (1/h)
fx_1	Fraction of inert COD generated by the death of the biomass
Q_r	Flow recycling between the clarifier and the reactor (1/h)
Q_w	Flow purge (1/h)
Y_h	Coefficient of performance of heterotrophic biomass
Q_L	Air flow in the aeration tank (1/h)
C_s	Constant saturation of dissolved oxygen (mg/l)
V_r	Volume of aeration basin (s)
V_{dec}	Volume of the settler (s)

μ_{max} Maximum growth rate of heterotrophic microorganisms (1/h)
K_s Coefficient half-saturation of readily biodegradable substrate for heterotrophic biomass (mg/l)
K_h Maximum specific rate for hydrolysis (1/h)
K_x Coefficient of half-saturation for hydrolysis of slowly biodegradable substrate

41.2 Appendix B: Constant Values

$S_{I,In} = 30,\ \ S_{S,In} = 50,\ \ X_{I,In} = 25,\ \ X_{S,In} = 125,\ \ V_r = 2,000\ \ V_{dec} = 1,500,$
$Y_H = 0.67,\ K_S = 20,\ K_H = 3,\ K_X = 0.03\ X_{S,In} = 125,\ X_{H,In} = 30,\ \mu_{max} = 0.67,$
$b_H = 0.62,\ C_S = 10,\ f_{x_1} = 0.086$

References

1. Henze M, Harremoes P, Jansen la Cour J, Arvin E (1995) Wastewater treatment, biological and chemical processes. Springer, Berlin
2. Olsson G, Newell B (1999) Wastewater treatment systems. IWA Publishing, London
3. Samuelsson P (2001) Modeling and control of activated sludge processes with nitrogen removal. Licentiate thesis, Uppsala University, Department of Systems and Control, IT Licentiate thesis
4. Mahmoud MS (2010) Coordinated control of waste water treatment process. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, WCE 2010, 30 June–2 July, London, UK, pp 2341–2346
5. Fragkoulis D, Roux G, Dahhou B (2007) Actuator fault isolation strategy to a waste water treatment process. In: Conference on systems and control, Mai 16–18, Marrakech, Morocco
6. Mahmoud MS, Hassan MF, Darwish MG (1985) Large scale control systems: theories and techniques. Marcel Dekker Inc, New York
7. Anderson BDO, Moore JB (1971) Linear optimal control. Prentice-Hall, Inc, Englewood Cliffs
8. Mahmoud MS, Hassan MF (1986) A decentralized water quality control scheme. IEEE Trans Syst Man Cybern 16:694–702
9. Costa C, Rodríguez J, Carmen Márquez M (2009) A simplified dynamic model for the activated sludge process with high strength wastewaters. Environ Model Assess 14:739–747

Chapter 42
Identical Parallel-Machine Scheduling and Worker Assignment Problem Using Genetic Algorithms to Minimize Makespan

Imran Ali Chaudhry and Sultan Mahmood

Abstract Identical parallel machine scheduling problem for minimizing the makespan is a very important production scheduling problem which has been proven to be NP-hard. The problem further compounds with additional constraints. Genetic algorithms (GA) have shown great advantages in solving the combinatorial optimization problem in view of its characteristic that has high efficiency and that is fit for practical application. In this chapter we present a spreadsheet based GA approach for minimizing the makespan for scheduling of a set of tasks for identical parallel machines and worker assignment to machines. The results obtained from the proposed approach are compared with two sets of benchmark problems consisting of 100 problems each. It has been demonstrated that the performance of proposed approach is superior to the results that have been obtained earlier. The proposed approach produces optimal solution for almost 95% of the problems demonstrating the effectiveness of the proposed approach. An empirical analysis of GA parameters has also been carried out to see the effect on the performance of the proposed algorithm.

42.1 Introduction

Scheduling is an important process widely used in manufacturing, production, management, computer science, and among others. There are many kinds of scheduling problems in the literature. One of them is the parallel machine

I. A. Chaudhry (✉) · S. Mahmood
College of Aeronautical Engineering, National University of Sciences and Technology, Risalpur, Pakistan
e-mail: imran_chaudhry@yahoo.com

S. Mahmood
e-mail: sultanrandhawa@hotmail.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_42,
© Springer Science+Business Media B.V. 2011

scheduling (PMS) problem, in which scheduling jobs in parallel machines is considered. Many real life problems can be modeled as PMS ones. On production lines, it is common to find more than one machine of each kind carrying out the production tasks. Other examples are docks—ships, teachers student groups, hospital assistance—patients, etc.

Pinedo [1] pointed out that a bank of machines in parallel is a situation that is important from both a theoretical and a practical point of view. From a theoretical point of view, it is a generalization of the single machine and a special case of the flexible flowshop. From a practical point of view, it is important because the occurrence of resources in parallel is common in the real world. Also, techniques for studying machines in parallel are often used in decomposition procedures for multistage systems. PMS comes down to assigning each operation to one of the machines and sequencing the operations assigned to the same machine.

In particular, the makespan becomes an objective of considerable interest when dealing with machines in parallel. In practice, one often has to deal with the problem of balancing the load on machines in parallel; by minimizing the makespan, the scheduler ensures a good balance. Scheduling jobs in identical machines for minimizing the makespan was proved as NP-hard [2, 3]. Thus, it is unlikely to obtain the optimal schedule through polynomial-time-bounded algorithms.

In this chapter we present a spreadsheet based GA approach to minimize the makespan for scheduling a set of tasks on identical parallel machines and worker assignment to the machines.

42.2 Literature Review

The first work on PMS problems started at the end of the 1950s with McNaughton [4] and then with Hu [5]. These types of problems have received continuous interest from researchers since then, due to their relevance to computer systems and production systems and thus the literature continues to increase. Various constraints have been taken into account and different criteria have been studied. For a detailed review of new trends in PMS see Lam and Xing [6] while an extensive survey has been given by Cheng & Sin [7] and Moktoff [8].

Makespan minimization is one of the most frequently studied criteria in the scheduling literature and also in identical parallel machine scheduling. Some of the recent papers reviewed in the subsequent paragraphs are restricted to this criterion.

Min and Cheng [9] present a kind of GA based on machine code for minimizing the makespan in identical parallel machine scheduling. Hey demonstrate that the proposed GA is efficient ad fit for large scale problems and has advantage over heuristic procedure and simulated annealing method.

Yalaoui and Chu [10] consider real-life identical PMS problem with sequence-dependent setup times and job splitting to minimize makespan and propose a

heuristic to solve this problem. Lee et al. [11] propose a simulated annealing method to generate near optimal solutions for the minimization of makespan in identical parallel machines. With the help of computational analysis they demonstrate that the proposed method is very accurate and outperforms the existing method. Akyol [12] also consider minimization of makespan in identical PMS and propose a Hopfield-like network that uses time-varying penalty parameters starting from zero and increases in a stepwise manner during iterations to overcome the tradeoff problem of the penalty function method.

Iori and Martello [13] consider scatter search algorithms and provide insights in the development of this heuristic technique and discuss the combinatorial difficulties of the problems through the analysis of extensive computational results. Akyol and Byhan [14] use neural networks for minimizing the maximum completion time (makespan) of jobs on identical parallel machines while Mokotoff et al. [15] propose algorithms that are based on list scheduling procedures.

Few other authors have considered other objectives along with minimization of makespan. Mohri et al. [16] consider minimizing two criteria, maximum completion time and maximum lateness for two and three identical PMS problems. Gupta and Ho [17] consider minimizing flowtime subject to optimal makespan on two identical parallel machines while Gupta and Ho [18] consider minimizing makespan subject to minimum flowtime on two identical parallel machines. Sricharoentham [19] consider memtic algorithm based on genetic algorithms for multiple objective scheduling in parallel machines.

Gupta et al. [20] consider makespan minimization on identical parallel machines subject to minimum total flow-time. Hong et al. [21] consider minimizing makespan of identical-machines scheduling problems with mold constraints. In this kind of problems, jobs are non-preemptive with mold constraints and several identical machines are available. They propose a GA based approach to solve this kind of problem. Ho et al. [22] present a two phase non-linear integer programming formulation for minimizing total weighted flowtime subject to minimum makespan on two identical parallel machines.

The papers reviewed above do not cater for the number of workers who are performing a certain task. Worker assignment scheduling problem has also been studied in the literature. In the classic PMS problem, no matter how many machines are involved, the number of workers at each machine may be ignored or assumed to be fixed and not taken into consideration. However, assigning more workers to work on the same job will decrease job completion time. Therefore, ignoring worker assignment decision may cause a managerial problem.

Hu [23, 24] considered the parallel machines models with decisions for job scheduling and worker assignment to minimize total tardiness and total flow time, respectively. A shortest processing time (SPT) heuristic and a largest marginal contribution (LMC) procedure were used to solve the job scheduling and worker assignment problems, respectively. The performance of these heuristics in Hu [23, 24] was further studied by Hu [25, 26], who concluded that these heuristics generate the same results no matter what the value of W (number of workers).

Chaudhry and Drake [27] also considered the minimization of total tardiness for the machine scheduling and worker assignment problems in identical parallel machines using GA. While Chaudhry [28] considered the minimization of total flow time for the worker assignment problem in identical parallel machine models using GA. Chaudhry et al. [29] proposed a GA to minimize makespan for machine scheduling and worker assignment problem in identical parallel machine models.

42.3 Problem Definition and Assumptions

In classical PMS problem, there are two essential issues to be dealt:

1. Partition jobs to machines.
2. Sequence jobs for each machine.

However, in the present research, the worker assignment scheduling problem needs to solve two sub-problems: how to assign jobs to machines and workers to machines? The objective is to minimize the makespan, which is defined as the total completion time of all the jobs. The performance of GA is compared with SPT/L(east)PA and L(argest)PA heuristics heuristic approach by Hu [30].

We assume that all machines are identical such that the processing time of a job is independent of machine. Deterministic processing times and due dates are assumed. Machine setup times are included in the processing time. The only difference between worker assignment scheduling problem and the classic scheduling problem is that the former has an additional constraint of worker assignment also. The processing times of the jobs are therefore dependent on the number of workers assigned to work on a particular machine.

The following notation is used to define the problem.

A_i, B_i, E_i	Integers, follow uniform distribution, such that $0 \leq A_i < 10$, $0 \leq B_i \leq 50$, $0 \leq E_i \leq 10$ (1st data set) $0 \leq A_i < 10$, $0 \leq B_i \leq 800$, $0 \leq E_i \leq 10$ (2nd data set)
n	Number of jobs
m	The number of parallel machine in the shop
p_i	The process time of job i
W	Number of workers in the shop
Wj	Number of workers assigned to machine m_j

We have to assign n jobs to m parallel machines taking into account the following characteristics:

1. Each job has only one operation.
2. The jobs have different processing times which depend on the number of workers assigned to that machine.
3. No job pre-emption/splitting is allowed.
4. Each job has its own due date.

5. The selected objective is to minimize the total tardiness.
6. Any machine can process any job.
7. Machine setup times are negligible.
8. No machine may process more than one job at a time.
9. Transportation time between the machines is negligible.
10. Number of jobs and machines are fixed.
11. There is a group of W workers that have the same abilities to perform the duties assigned to them.
12. All m machines, n jobs and W workers are available at time zero.
13. Assume processing time function has a simplified form of $p_i(W_j) = A_i + B_i/(E_i \times W_j)$ where $p_i(W_j)$ is the processing time of job I that is processed on machine j in which the number of W_j workers are assigned on it, A_i is a fixed constant and not affected by the number of workers, $B_i/(E_i \times W_j)$ is the variable part and affected by the number of workers.
14. The number of workers assigned to each machine needs to be decided before any job can be processed and they will not be re-assigned until all the jobs have been completed.

The simulation experiments have been carried out for 12 jobs to be scheduled on three parallel machines with 10 workers.

42.4 Genetic Algorithms

GAs is one of problem solving systems based on the principles of evolution and hereditary, each system start with an initial set of random solutions and use a process similar to biological evolution to improve upon them that encourage the survival of the fittest. The best overall solution becomes the candidate solution to the problem. A detailed introduction to Gas is given in Goldberg [31]. The earliest application of GA has been reported by Davis [31]. For a recent review of GA application in scheduling is given in Chaudhry and Drake [27] and Chaudhry [28].

In this study, the tool use for carrying out the GAs is a commercial software package called EVOLVER™ [32], which functions as an add-into Microsoft Excel™. Evolver has been used for the purpose of production scheduling recently by number of researchers [27–29, 33–35].

The objective function, variables (adjustable cells), and the constraints are readily specified by highlighting the corresponding spreadsheet cells. The scheduler/evaluator portion of the model is constructed by using the spreadsheet's built in function. The schematic in Fig. 42.1 illustrates the integration of the GA with the spreadsheet.

The advantage of the proposed method is that the program runs in the background freeing the user to work in the foreground. When the program finds the best result it notifies the user and places the values into the spreadsheet for analysis. This is an excellent design strategy given the importance of interfacing with spreadsheet in business.

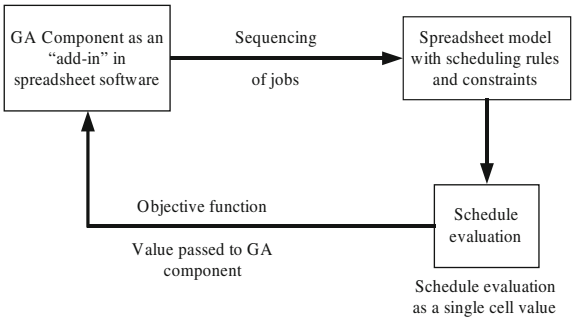

GA Component as an "add-in" in spreadsheet software
Sequencing
of jobs
Spreadsheet model with scheduling rules and constraints
Objective function
Value passed to GA component
Schedule evaluation
Schedule evaluation as a single cell value

Fig. 42.1 Integration of GA and spreadsheet

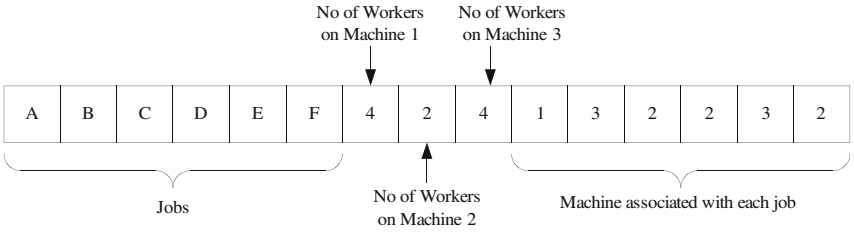


Fig. 42.2 Chromosome representation

42.4.1 Chromosome Representation

The chromosome representation for the machine scheduling and worker assignment problem is shown in the Fig. 42.2.

For determining the ordering of the jobs i.e., for the first six genes we needed to handle the permutation representation. The next three genes represent the number of workers assigned to each machine stating that four workers are assigned to machine 1, two workers to machine 2 and four workers to machine 3. While the machine assignment block can be read as the assignment of jobs to each of the six machines, whereby the assignment of the jobs to machines would be as follows: Job A to be processed on machine 1; job B on machine 3, job C on machine 2, job D on machine 2, job E on machine 3 and job F on machine 2, respectively.

Each of the block in Fig. 42.2 is calculated at a different location in the spreadsheet which is in turn linked to the calculation of the objective function i.e., the makespan.

42.4.2 Reproduction/Selection

Evolver uses a steady state approach. This means that only one organism rather than an entire population is replaced at a time. The number of generations can be found by dividing the trials by the size of the population. As far as parent selection

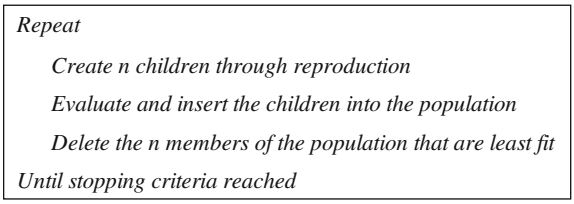

Repeat
Create n children through reproduction
Evaluate and insert the children into the population
Delete the n members of the population that are least fit
Until stopping criteria reached

Fig. 42.3 Steady state algorithm

is concerned, in Evolver, parents are chosen using a rank-based mechanism. This procedure begins by rank ordering the population by fitness. Next an assignment function gives each individual a probability of inclusion. The assignment function can be linear or nonlinear. A Roulette wheel is then built with the slots determined by the assignment function. The next generation of an n-sized population is built by giving the wheel n spins. This procedure guides selection towards the better performing members of the population but does not force any particular individual into the next generation. Figure 42.3 describes the steady state algorithm.

42.4.3 Crossover Operator

As for the first six genes of the chromosome, we needed to handle permutation representation; the "order solving method" of Evolver was used. This method applies order crossover operator [36]. This selects items randomly from one parent, finds their place in the other parent, and copies the remaining items into the second parent in the same order as they appear in the first parent. This preserves some of the sub-orderings in the original parents while creating some new sub-orderings. Figure 42.4 shows the crossover operator as described above.

For the number of workers and machine assignment the "recipe solving method" of Evolver was used. The "recipe solving method" implements uniform crossover [36]. This means that instead of chopping the list of variables in a given scenario at some point and dealing with each of the two blocks (called "single-point" or "double-point" crossover), two groups are formed by randomly selecting items to be in one group or another. Traditional x-point crossovers may bias the search with the irrelevant position of the variables, whereas the uniform crossover method is considered better at preserving schema, and can generate any schema from the two parents. For worker assignment (genes 7–9 in Fig. 42.2) and machine assignment (genes 10–15 in Fig. 42.2), we have a set of variables that are to be adjusted and can be varied independently of one other. In the spreadsheet model presented in this chapter, the constraint placed on the workers and machines is to set the range that the variable must fall between. Similarly, as there are ten workers available in the shop, another constraint that ensures that only valid solutions are retained in the population is that the sum of values generated for genes 7–9 should be equal to 10. This ensures that no invalid solution is retained in the population pool. Figure 42.5 shows a typical uniform crossover operator.

Fig. 42.4 Order crossover operator

Position:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Parent 1 (P1):	1	2	3	4	5	6	7	8	9
Binary Template:	*0*	*1*	*1*	*0*	*1*	*1*	*0*	*0*	*1*
Parent 2 (P2):	4	9	5	8	3	6	7	1	2
Offspring (O):	4	2	3	8	5	6	7	1	9

Fig. 42.5 Uniform crossover—a given % of the organism is randomly selected

42.4.4 Mutation Operator

To preserve all the original values, the "order solving method" performs mutation by swapping the positions of some variables in the organism. The number of swaps performed is increased or decreased proportionately to the increase and decrease of the mutation rate setting (from 0 to 1).

The "recipe solving method" performs mutation by looking at each variable individually. A random number between 0 and 1 is generated for each of the variables in the organism, and if a variable gets a number that is less than or equal to the mutation rate (for example, 0.06), then that variable is mutated. The amount and nature of the mutation is automatically determined by a proprietary algorithm. Mutating a variable involves replacing it with a randomly generated value (within its valid min–max range).

42.5 Experimental Results

42.5.1 Implementation Details

The job order (first six genes) in Fig. 42.2 is a permutation of a list of jobs where we are trying to find the best way to arrange a set of given jobs. This permutation is independent of the number of workers on each machine and assignment of job to a particular machine. However, the objective function is calculated keeping in view all the constraints which are discussed in the next paragraph.

For number of workers on each machine and the machine corresponding to each job, i.e., for genes 7–9 and 10–15 (Fig. 42.2), random integer numbers are generated by the GA subject to the constraints that have been defined in the initial setup of the GA. Constraints are basically the conditions that must be met for a solution to be valid. The constraint imposed on the number of workers on each machine is that the sum of all workers assigned to the three machines should always be 10, which is the total number of workers available in the shop. The range for random integer is from 1 to 10. Therefore, only those solutions are kept

in the population where the sum of all workers is 10. Hence, only integer values between 1 and 10 (both limits inclusive) are generated, while the constraint for machine corresponding to each job is that an integer number is selected from among 1, 2, and 3 (which are machine numbers).

42.5.2 Computational Analysis

In order to check the effectiveness of the proposed spreadsheet based GA approach two data set of 100 problems each [30] were used. The problems in both the data sets have 12 jobs to be scheduled on three identical parallel machines where the number of workers available is equal to 10. The only difference among the two data sets in terms of the value of the variable B_i, for first data set it is $0 \leq B_i \leq 50$ and for second data set it is $0 \leq B_i \leq 800$. The problems have been simulated on a P1V 1.7 GHz computer having 512 MB RAM. By conducting repeated tests, we found that the best values to be set for the number of population, the crossover rate, and the mutation rate are 60, 0.65, and 0.09, respectively. Therefore, for each of the run, same set of parameter setting as described previously have been used, which correspond to 3 min 20 s on a PIV 1.7 GHz computer having 512 MB RAM.

Tables 42.1 and 42.2 give a summary of the results for the first and second data sets, respectively. The summary shows the performance of proposed GA approach with the SPT/L(east)PA and L(argest)PA heuristic proposed by Hu [30].

GA produced superior solution as compared to the SPT/L(east)PA and L(argest)PA heuristic. For the first data set GA found 95 optimal solutions as compared to the heuristic which found optimal solution for 22 problems. While for second data set these values were 96 and 27, respectively. The average time for the GA for first and second data set was to find the best solution was 70 and 86 s, respectively. As compared to the optimal solution, for the first data set the maximum percentage error was 7.87 and 65.47% for the GA and SPT/L(east)PA and L(argest)PA heuristic [30], respectively. While for the second data set these maximum percentage errors were 8.94 and 38.31%, respectively.

As compared to SPT/L(east)PA and L(argest)PA heuristic, for the first data set GA produced same solution for 22 problems and better for 77 problems. For the

Table 42.1 Summary of results for the two approaches—1st data set

		GA	SPT/L(east)PA and L(argest)PA heuristic [30]
# of problems simulated		100	100
# of optimal solution obtained		95	22
Max percentage error (%)		7.87	65.47
Avg Time to find best solution		70 s	–
Comparison of GA with other methods	Same	–	22
	Better	–	77
	Worse	–	1

Table 42.2 Summary of results for the two approaches—2nd data set

		GA	SPT/L(east)PA and L(argest)PA heuristic [30]
# of problems simulated		100	100
# of optimal solution obtained		96	18
Max percentage error (%)		8.94	38.31
Avg Time to find best solution		86 s	–
Comparison of GA with other methods	Same	–	27
	Better	–	71
	Worse	–	2

Table 42.3 Combinations of GA parameter levels

Population size	20, 40, 60, 80
Mutation rate	0.01, 0.03, 0.05, 0.07, 0.09
Crossover rate	0.2, 0.4, 0.6, 0.8

second data set these were 27 and 71, respectively. While for only one problem in first data set and two problems in second data set GA produced worse results as compared to SPT/L(east)PA and L(argest)PA heuristic.

42.6 Empirical Analysis of the Effect of GA Parameters

The performance of a GA depends upon population size, crossover, and mutation rate. Detailed experiments have been carried out to study the effect of these three parameters. The application of the GA to the schedule optimization is repeated here for each of the combinations of the parameter levels given in Table 42.3 for both the data sets mentioned in Sect. 42.5.2.

For each combination of the parameters, the GA is run for ten different random initial populations. These ten populations are different for each combination. Thus, in total, the GA is run 800 times. The values chosen for the population size are representative of the range of values typically seen in the literature, with 0.09 being included to highlight the effect of a relatively large mutation rate. The crossover rates chosen are representative of the entire range.

The results show that the performance of the GA is insensitive to the crossover rate. There is also a degree of insensitivity to the mutation rate, which possessed a 'range' over which its value is suitable rather than a more precise value. Furthermore, performance is fairly insensitive to population size, provided the mutation rate is in the 'good' range. The various set of problems that have been solved in this chapter demonstrate that GA-spreadsheet approach presented in this chapter performs well on a wide range of shop models with a 'general purpose' set of parameters.

42.7 Conclusions

This chapter presented a spreadsheet based general purpose GA solution methodology for the scheduling of set of job on parallel machines and worker assignment to machines. The spreadsheet GA implementation has been found to be easy to implement catering for the peculiarities of any environment. Moreover, the spreadsheet environment makes it very suitable to carry out what if analysis. The results in Tables 42.1 and 42.2 clearly show the superiority of proposed GA approach as compared to an earlier study by Hu [29]. The spreadsheet model can be easily customized to include additional jobs, machines or workers without actually changing the logic of the GA routine thus making it a general purpose scheduling approach.

The key advantage of GAs portrayed here is that they provide a general purpose solution to the scheduling problem which is not problem-specific, with the peculiarities of any particular scenario being accounted for in fitness function without disturbing the logic of the standard optimization routine. The GA can be combined with a rule set to eliminate undesirable schedules by capturing the expertise of the human scheduler.

The empirical analysis of the GA parameters shows that crossover rate is often an insignificant factor in the performance of the GA. There is a degree of insensitivity to the mutation rate in so far as the 'good' values form a fairly broad range rather than coming at a more precise point. However, outside of the 'good' range performance soon deteriorates greatly. It is observed that a low mutation rate is desirable when the population size is large and a higher mutation rate is desirable when the population size is small. The sensitivity to population size is greatly reduced when the mutation rate is in the 'good' range. As a consequence of these findings, it has been argued that GAs have the potential to make a general-purpose real-world scheduler.

References

1. Pinedo ML (2008) Scheduling theory algorithms and systems. Springer Science, New York
2. Brucker P (1995) Scheduling algorithms. Springer, New York
3. Garey MR, Johnson DS (1979) Computers and intractability a guide to the theory of NP-completeness. Freeman, San Francisco
4. McNaugton R (1959) Scheduling with deadlines and loss functions. Manag Sci 6:1–12
5. Hu TC (1961) Parallel sequencing and assembly line problems. Oper Res 9:841–948
6. Lam K, Xing W (1997) New trends in parallel machine scheduling. Int J Oper Prod Man 17(3):326–338
7. Cheng T, Sin C (1990) A state-of-the-art review of parallel machine scheduling research. Eur J Oper Res 47:271–292
8. Mokotoff E (2001) Parallel machine scheduling problems: a survey. Asia Pac J Oper Res 18:193–242
9. Min L, Cheng W (1999) A genetic algorithm for minimizing the makespan in the case of scheduling identical parallel machines. Artif Intell Eng 13(4):399–403

10. Yalaoui F, Chu C (2003) An efficient heuristic approach for parallel machine scheduling with job splitting and sequence-dependent setup times. IIE Trans 35(2):183–190
11. Lee W-C, Wu C-C, Chen P (2006) A simulated annealing approach to makespan minimization on identical parallel machines. Int J Adv Manuf Tech 31:328–334
12. Akyol DE (2007) Identical parallel machine scheduling with dynamical networks using time-varying penalty parameters. In: Levner E (ed) Multiprocessor scheduling: theory applications. Itech Education and Publishing, Vienna, pp 293–314
13. Iori M, Martello S (2008) Scatter search algorithms for identical parallel machine scheduling problems. Stud Comp Intell 128:41–59
14. Akyol DE, Bayhan GM (2006) Minimizing makespan on identical parallel machines using neural networks. Lect Notes Comput Sci 4234:553–562
15. Mokotoff E, Jimeno JL, Gutiérrez AI (2001) List scheduling algorithms to minimize the makespan on identical parallel machines. TOP 9(2):243–269
16. Mohri S, Masuda T, Ishii H (1999) Bi-criteria scheduling problem on three identical parallel machines. Int J Prod Econ 60–61:529–536
17. Gupta JND, Ho JC (2000) Minimizing flowtime subject to optimal makespan on two identical parallel machines. Pesqui Oper 20(1):5–17
18. Gupta JND, Ho JC (2001) Minimizing makespan subject to minimum flowtime on two identical parallel machines. Comput Oper Res 28(7):705–717
19. Sricharoentham C (2003) Multiple-objective scheduling in parallel machines using memtic algorithm. MS Thesis Texas Tech University, USA
20. Gupta JND, Ho JC, Ruiz-Torres AJ (2004) Makespan minimization on identical parallel machines subject to minimum total flow-time. J Chin Inst Ind Eng 21(3):220–229
21. Hong TP, Sun PC, Jou SS (2009) Evolutionary computation for minimizing makespan on identical machines with mold constraints. WSEAS Trans Syst Control 7(4):339–348
22. Ho JC, López FJ, Ruiz-Torres AJ, Tseng TL (2009) Minimizing total weighted flow-time subject to minimum makespan on two identical parallel machines. J Intell Manuf doi:10.1007/s10845-009-0270-1
23. Hu PC (2004) Minimizing total tardiness for the worker assignment scheduling problem in identical parallel-machine models. Int J Adv Manuf Tech 23(5–6):383–388
24. Hu PC (2005) Minimizing total flow time for the worker assignment scheduling problem in the identical parallel-machine models. Int J Adv Manuf Tech 25(9–10):1046–1052
25. Hu PC (2006) Further study of minimizing total tardiness for the worker assignment scheduling problem in the identical parallel machine models. Int J Adv Manuf Tech 29:165–169
26. Hu PC (2006) Further study of minimizing total flowtime for the worker assignment scheduling problem in the identical parallel machine models. Int J Adv Manuf Tech 29(7–8):753–757
27. Chaudhry IA, Drake PR (2008) Minimizing total tardiness for the machine scheduling and worker assignment problems in identical parallel machines using genetic algorithms. Int J Adv Manuf Tech 42:581–594
28. Chaudhry IA (2010) Minimizing flow time for the worker assignment problem in identical parallel machine models using GA. Int J Adv Manuf Tech 48(5–8):747–760
29. Chaudhry IA, Mahmood S, Ahmad R (2010) Minimizing makespan for machine scheduling and worker assignment problem in identical parallel machine models using GA. Lecture notes in engineering and computer science: Proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July, London, UK, pp 2464–2469
30. Hu PC Online research data set available at http://sparc.nfu.edu.tw/~pchu/research_data.htm
31. Goldberg DE (1989) Genetic algorithms in search optimization and machine learning. Addison-Wesley, Boston
32. Palisade Corporation (1998) Evolver: the genetic algorithm super solver. New York, USA
33. Chaudhry IA, Drake PR (2008) Minimizing flow time variance in a single machine system using genetic algorithms. Int J Adv Manuf Tech 39:355–366

34. Hayat N, Wirth A (1997) Genetic algorithms and machine scheduling with class setups. Int J Comput Eng Manag 5(2):10–23
35. Ruiz R, Maroto C (2002) Flexible manufacturing in the ceramic tile industry. In: Proceedings of eighth international workshop on project management and scheduling, April 3–5, Valencia Spain
36. Davis L (1991) Handbook of genetic algorithms. Van Nostrand Reinhold, New York

Chapter 43
Dimensional Accuracy Achievable in Wire-Cut Electrical Discharge Machining

Mohammad Nazrul Islam, Noor Hakim Rafai and Sarmilan Santhosam Subramanian

Abstract Wire-cut electrical discharge machining (WEDM) is a popular choice for machining hard and difficult to machine materials with very close tolerances. However, the widely held assumption of the high accuracy of WEDM needs to be investigated, which is the primary aim of this research. This work presents the experimental and analytical results of an investigation into the dimensional accuracy achievable in WEDM. Three techniques—traditional analysis, the Taguchi method, and Pareto ANOVA—are employed to determine the effects of six major controllable machining parameters: the discharge current, pulse duration, pulse gap frequency, wire speed, wire tension, and dielectric flow rate on three key dimensional accuracy characteristics of the prismatic component parts—linear dimensional errors, flatness errors, and perpendicularity errors of corner surfaces. Subsequently, the input parameters are optimized in order to maximize the dimensional accuracy characteristics. The results indicate that the dimensional accuracy that is achievable in WEDM is not as high as anticipated.

M. N. Islam (✉)
Department of Mechanical Engineering, Curtin University, GPO Box U1987, Perth, WA 6845, Australia
e-mail: m.n.islam@curtin.edu.au

N. H. Rafai
Department of Manufacturing and Industrial Engineering, Universiti Tun Hussein Onn Malaysia (UTHM), 86400, Parit Raja, Batu Pahat, Johor, Malaysia
e-mail: nhakim@uthm.edu.my

S. S. Subramanian
Holcim (Australia) Pty Ltd., PO Box 138, Gosnells, WA 6990, Australia
e-mail: sarmilan.subramanian@holcim.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_43,
© Springer Science+Business Media B.V. 2011

43.1 Introduction

Wire-cut electrical discharge machining (WEDM) is one of the most widely used non-traditional machining processes in current manufacturing. It involves the removal of metal by discharging an electrical current from a pulsating DC power supply across a thin interelectrode gap between the tool and the workpiece. It is a popular choice for machining hard and difficult to machine materials with very close tolerances. Generally, WEDM is perceived to be an extremely accurate process and there are various reasons for this perception. Firstly, in WEDM, no direct contact takes place between the cutting tool (electrode) and the workpiece; as a result, the adverse effects—mechanical stresses, chatter, and vibration—normally present in traditional machining are eliminated. Secondly, the wire used as a cutting tool has high mechanical properties and small diameters (0.076–0.30 mm [1]), which is believed to produce very fine, precise, and clean cuts. Finally, in WEDM, the movements of the workpiece during cutting are controlled by a highly accurate computer numerical controlled (CNC) system (with positioning accuracy up to ± 0.5 μm [1]); as a result, the effects of positioning errors present in conventional machining are significantly diminished. However, this perception of the high accuracy of WEDM needs to be investigated, which the primary objective of this project.

Since its advent in the early 1970s, there have been numerous papers reported on various aspects of WEDM, such as metal removal rate [2, 3], surface finish [3–5], and process modeling [6]. However, there has been less interest in the dimensional accuracy achievable by this process [7–9]. In addition, the reported studies on WEDM concentrated on a single dimensional accuracy characteristic only and, as such, did not take into account their combined effects on machined parts. Therefore, in [10], an attempt was made to examine three key dimensional accuracy characteristics of parts produced by WEDM concurrently, and to find the optimum combination major controllable input parameters. The work presented here is an extended and revised version of our previous work [10].

43.2 Scope

The main objective of this project is to investigate the dimensional accuracy characteristics achievable of typical component parts produced by the WEDM process. For the sake of simplicity, in this study, a rectangular block is selected as a test part, details of which are given in the following section. For such a part, the three most important dimensional accuracy characteristics are: (i) linear dimensional error, (ii) the flatness of the surfaces produced, and (iii) the perpendicularity error of the corners. Thus, these characteristics were selected here to monitor the quality of the parts produced by WEDM. The six independent input parameters chosen are: (i) discharge current, (ii) pulse duration, (iii) pulse gap frequency, (iv) wire speed, (v) wire tension, and (vi) dielectric flow rate. A general purpose

coordinate machine (CMM) is employed for the measurement of the output parameters. The results are analyzed by three techniques: (i) traditional analysis, (ii) Pareto analysis of variation (ANOVA), and (iii) Taguchi's signal-to-noise ratio (*S/N*) analysis. The expected outcomes of this project are: (i) to get a clear picture of the machining accuracy achievable in WEDM, (ii) to find out the influences of the six input parameters on the accuracy of a typical component part produced by WEDM, and (iii) to optimize the input parameters.

In the traditional analysis, the mean values of the measured variables were used. For the Taguchi method, the *signal-to-noise ratio* was calculated using the following formula [11]:

$$S/N = -10 \log \frac{1}{n} \left(\sum_{i=1}^{n} \frac{1}{y_i^2} \right) \tag{43.1}$$

where *S/N* is the signal-to-noise ratio (in *dB*), n is the number of observations, and y is the observed data.

The above formula is suitable for quality characteristics in which "the smaller the better" holds true. This is the case for all three quality characteristics considered. The higher the value of the *S/N* ratio, the better the result is because it guarantees optimum quality with minimum variance. A thorough treatment of the Taguchi method can be found in [11]. Pareto ANOVA is a simplified ANOVA method that does not require an ANOVA table; further details of Pareto ANOVA can be found in [12].

43.3 Experimental Work

The experiments were planned using Taguchi's orthogonal array methodology and a three-level L_{27} (3^{13}) orthogonal array was selected for our experiments. A total of 27 experimental runs were conducted. Besides the six main effects (*A*–*F*), two interaction effects were also selected for analysis. The selected interactions are between the discharge current and pulse duration ($A \times B$) and between the discharge current and dielectric flow rate ($A \times F$).

Even though one of the main advantages of using WEDM is its ability to cut hard and difficult to machine materials with low machinability ratings, in this study, mild steel 1040 was chosen as the work material because of its low cost and availability. Nevertheless, it is anticipated that hard and difficult to machine materials will produce inferior machining accuracy. The designed sizes for the rectangular test part ($L \times W \times H$) were 20 × 10 × 15 mm. Cutting was performed on a 15 mm plate and the rectangular block was extracted from the plate by means of cutting along the contour. The height remains as it is because machining was not done on the height.

A total of 27 test parts marked TP1–TP27 were produced on a FANUC ROBOCUT α oiD, manufactured by FANUC, Japan. It is a high performance wire-cut EDM equipped with digital servo technology. The available machining space for this machine is 370 × 270 × 255 mm along the *X*-, *Y*-, and *Z*-axes, respectively.

Table 43.1 Input variables

Input parameters	Unit	Symbol	Levels		
			Level 0	Level 1	Level 2
Discharge current	amp	A	16.00	20.00	24.00
Pulse duration	ms	B	3.00	6.00	9.00
Pulse gap frequency	kHz	C	40.00	50.00	60.00
Wire speed	m/min	D	7.00	8.00	9.00
Wire tension	g	E	1000	1150	1300
Dielectric flow rate	MPa	F	0.14	0.20	0.26

The wire used is an EDM brass wire with a 0.25 mm diameter, well-known for its excellent mechanical properties and its capability to achieve high dimensional accuracy.

The six most important input variables were selected after an extensive literature review and subsequent preliminary investigations. Their limits were set on the basis of the capacity and limiting cutting conditions of the WEDM, ensuring continuous cutting by avoiding the breakage of the wire; details are given in Table 43.1.

The precision measurements were taken by a Discovery Model D-8 coordinate measuring machine (CMM), manufactured by Sheffield, UK. The probes used were spherical probes with a star configuration, manufactured by Renishaw Electrical Ltd. The linear size of the test parts was calculated using the standard built-in software package of the CMM. For each length feature, 14 measurements were taken at a 1 mm height step. The difference between the measured size and the designed size is the linear dimensional error, thus, a positive error indicates over sizing of a feature. A large number of points, 5 × 14 on the long faces and 3 × 14 on the short faces, respectively, were measured to determine the flatness error and to monitor the surface profile at different cross-sections. Additional measurements were taken at three different heights to determine the perpendicularity error of each corner angle. A positive perpendicularity error indicates that the corner angle is larger than 90^{o}.

43.4 Results and Analysis

An enormous amount of data was obtained and subsequently analyzed. Due to space constraints, only a few are illustrated, although in the analysis of the work, all these relationships were considered at different stages.

43.4.1 Linear Dimensional Errors

The results of the linear dimensional errors are shown in Table 43.2. It is noted that in all cases, the measured mean linear dimension size is less than the designed size. This indicates that the test parts have been overcut. Overcutting is a common problem in WEDM [13]. The main reason behind this is that during WEDM operation, the size of the cavity created in the workpiece is larger than the wire

Table 43.2 Linear dimensional error results

Input parameters	Unit	WEDM		End milling [18]	
		Length	Width	Length	Width
Design size	mm	20.000	10.000	200.000	75.000
Measured mean size	mm	19.787	9.902	199.966	74.963
Linear dimensional error	μm	−213	−98	−34	−37
Range of measurement	μm	97	193	36	35
6× Standard deviation	μm	146	136	51	53
Calculated IT grade		11.352	11.713	7.277	8.146

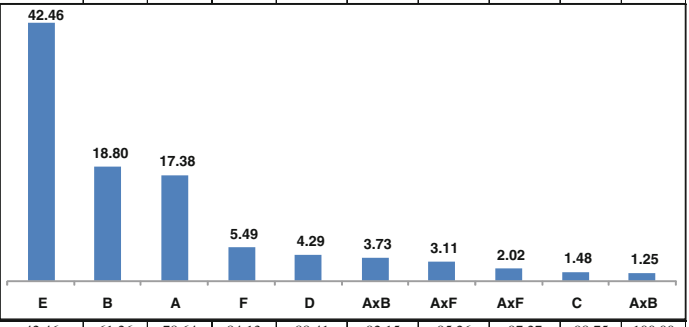
diameter. The exact size of the overcut is difficult to predict, but it is known to be proportional to the discharge current [13]. This explains the higher contributing effect of the discharge current (A) on the linear dimensional errors shown in the Pareto ANOVA (Table 43.3). While most of the modern WEDMs are equipped with inbuilt overcut error compensation means, it appears that those measures were not enough to overcome this problem.

The international tolerance (IT) grade is often used as a measure to represent the precision of a machining process, where the higher is the IT grade number and the lower is the precision of a process. The following formula has been utilized by several authors [14–16] to estimate the process capability tolerance achievable through various manufacturing processes:

$$\mathrm{PC} = (0.45\sqrt[3]{X} + 0.001X)10^{\frac{\mathrm{IT}-16}{5}} \tag{43.2}$$

where PC is the process capability tolerance (mm), X is the manufactured dimension (mm), and IT is the International Tolerance grade number.

Table 43.3 Pareto ANOVA for linear dimensional error

Sum at factor level	Factor and interaction									
	A	B	AxB	AxB	F	AxF	AxF	C	D	E
0	300.99	316.83	298.50	295.52	299.02	306.48	295.56	297.84	307.38	264.938
1	313.46	292.76	307.02	301.62	292.37	294.78	305.09	297.16	298.82	297.008
2	285.18	290.04	294.10	302.50	308.23	298.37	298.99	304.63	293.43	307.442
Sum of sq. of difference (S)	1204.71	1303.73	258.96	86.65	380.56	215.84	139.75	102.32	297.14	2943.92
Contribution ratio (%)	17.38	18.80	3.73	1.25	5.49	3.11	2.02	1.48	4.29	42.46
										
Cumulative contribution	42.46	61.26	78.64	84.13	88.41	92.15	95.26	97.27	98.75	100.00
Check on significant interaction			AxB two-way table							
Optimum combination of significant factor level			A1 B0 C2 D2 E0 F2							

The expected IT grades for WEDM and end milling are calculated applying Eq. 43.2, where six times standard division values shown in Table 43.2 represent the process capability tolerances. The calculated values demonstrate that in terms of linear dimensional accuracy, the WEDM performed poorly and its precision level is far less than CNC end milling.

The Pareto ANOVA for linear dimensional errors given in Table 43.3 illustrates that wire tension (E) has the most significant effect on linear dimensional errors ($P = 42.46\%$). The wire tension influences dimensional errors by a phenomenon known as *wire lag*, caused by the static deflection of the wire electrode. The effect of the wire lag on surface errors is discussed in the following subsection. The two other major contributing factors to linear dimensional errors are: pulse duration (B) ($P = 18.80\%$) and discharge rate (A) ($P = 17.38\%$). It is worth pointing out that the total of all of the individual effects ($P \cong 90\%$) is much higher than the total of all the interaction effects (P $\cong$ 10%). Therefore, it will be relatively easy to control the linear dimensional error through proper selection of the independent input parameters.

The response graphs for the dimensional errors are shown in Fig. 43.1a. Based on the *S/N* ratio and Pareto ANOVA, it was found that the combination for achieving a low linear dimensional error value was $A_1B_0C_2D_2E_0F_2$; that is, a medium discharge current, low pulse rate, high pulse gap frequency, high wire speed, low wire tension, and high dielectric flow rate.

43.4.2 Flatness Errors

For prismatic components, a flatness error is another important quality characteristic, which is geometric in nature. It is particularly important for parts where mating takes place across a surface area in an air-tight or liquid-tight manner. The flatness tolerances are also applied on all *principle datum surfaces* to ensure the integrity of measurement. *Flatness* is the condition of a surface having all elements in one plane [17]. A flatness error specifies a zone defined by two parallel planes between which the entire surface must lie.

The flatness error results given in Table 43.4 illustrate that the surfaces produced by WEDM have flatness errors about ten times higher compared to surfaces produced by CNC end milling.

It is worth noting that the flatness data does not give any indication of the shape of the cross-sectional profile. Therefore, in this study, in addition to flatness data, the cross-sectional profile of the test parts are monitored, which may help in understanding the flatness error-forming mechanics. A typical surface profile created by WEDM is depicted in Fig. 43.2, where $z = 0$ represents the bottom of the cut. Similar drum-shaped surface profiles have been observed in [8], which are believed to be caused by wire bending and vibration.

It is worth pointing out that in a vertical plane, the surface errors for WEDM and CNC end milling are comparable, however, the main source for the high

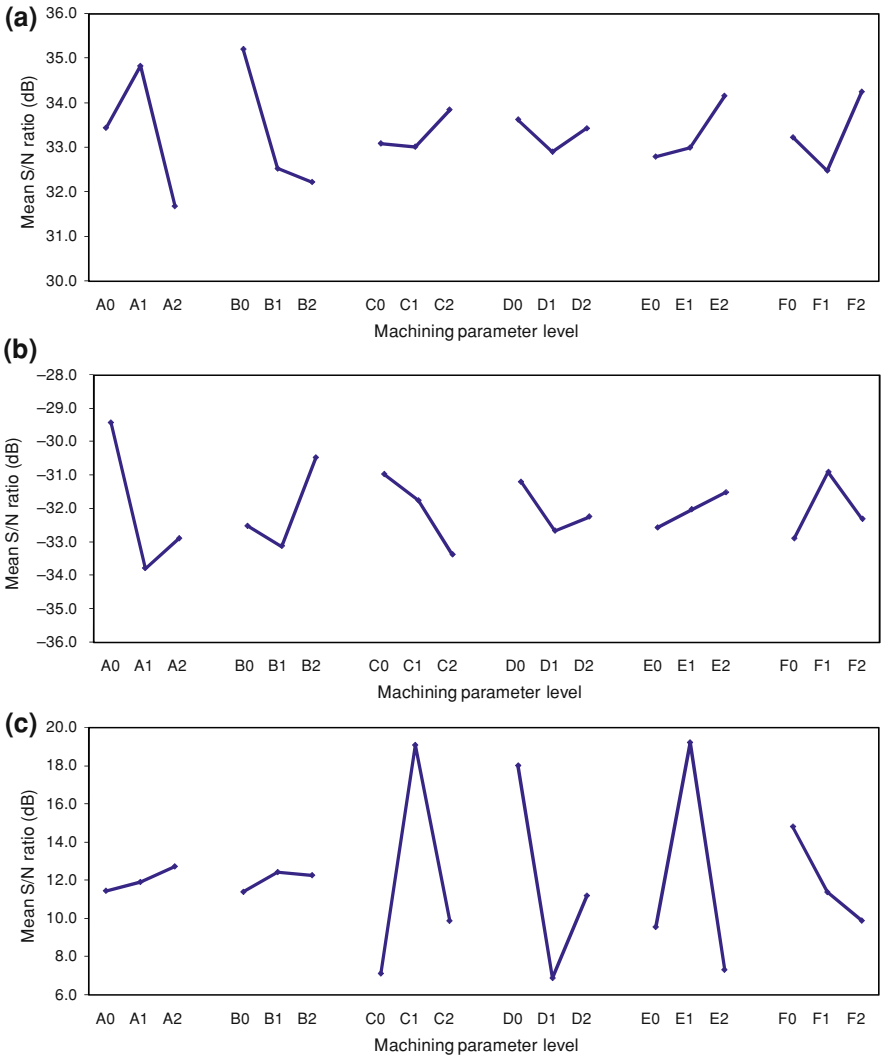


Fig. 43.1 Response graphs for WEDM: **a** linear dimensional error, **b** flatness error and **c** perpendicularity error

Table 43.4 Flatness error results

Input parameters	Unit	WEDM	End milling [18]
Feature size ($L \times H$)	mm	20 × 15	200 × 12
Measure mean flatness error	μm	48	17
Range of measurement	μm	189	19
6× Standard deviation	μm	271	28

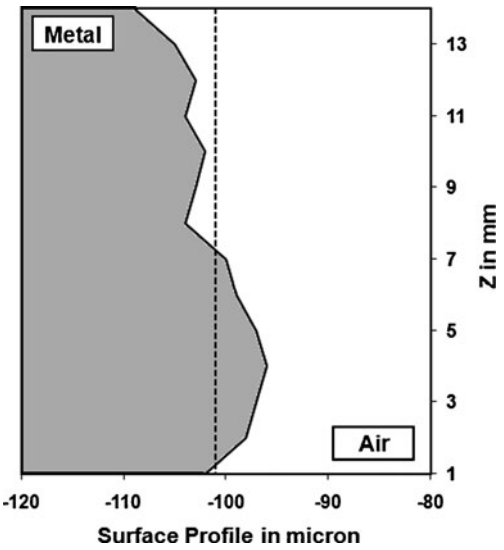


Fig. 43.2 A typical surface profile created by WEDM

flatness error values in WEDM were caused by the errors at the corners. The problem of erosion of the corner shapes has been identified by a number of researchers [7–9], and is also a result of the wire lag phenomenon. The combined effects of erosion of corners and the drum-shaped surfaces produced resulted in high flatness errors for the surfaces produced by WEDM.

The Pareto ANOVA shown in Table 43.5 illustrates that the most significant impendent parameter affecting the flatness error was the discharge current (A) ($P = 30.23\%$), followed by wire tension (E) ($P = 15.70\%$) and pulse duration

Table 43.5 Pareto ANOVA for flatness error

Sum at factor level	Factor and interaction									
	A	B	AxB	AxB	F	AxF	AxF	C	D	E
0	-264.80	-292.62	-285.41	-291.40	-295.97	-299.25	-274.28	-278.68	-293.02	-260.311
1	-304.02	-298.00	-280.99	-285.56	-278.09	-285.10	-299.41	-285.80	-292.02	-288.134
2	-295.97	-274.16	-298.38	-287.83	-290.72	-280.44	-291.09	-300.31	-279.74	-283.603
Sum of sq. of difference (S)	2574.54	938.00	490.29	51.98	507.09	575.88	982.81	729.44	328.07	1337.15
Contribution ratio (%)	30.23	11.02	5.76	0.61	5.96	6.76	11.54	8.57	3.85	15.70
	30.23	15.70	11.54	11.02	8.57	6.76	5.96	5.76	3.85	0.61
	A	E	AxF	B	C	AxF	F	AxB	D	AxB
Cumulative contribution	30.23	45.94	57.48	68.49	77.06	83.82	89.78	95.54	99.39	100.00
Check on significant interaction			AxF two-way table							
Optimum combination of significant factor level			A0 B2 C0 D0 E2 F1							

Table 43.6 Perpendicularity error results

Input parameters	Unit	WEDM	End milling [18]
Feature size ($L \times W \times H$)	mm	$20 \times 10 \times 15$	$200 \times 45 \times 12$
Measure mean perpendicularity error	deg	−0.524	0.072
Range of measurement	deg	2.766	0.527
6× Standard deviation	deg	4.317	1.089

(B) ($P = 11.02\%$). The total of all interaction effects on flatness errors is relatively higher ($P \cong 24.7\%$) compared to the total of all interaction effects on dimensional errors ($P \cong 10\%$). Therefore, it will be more difficult to control flatness errors through the individual selection of input parameters. The response graphs for flatness errors are shown in Fig. 43.1b. Based on the *S*/*N* ratio and Pareto ANOVA, it was found that the combination for achieving a low linear dimensional error value was $A_0B_2C_0D_0E_2F_1$; that is, a low discharge current, high pulse rate, low pulse gap frequency, low wire speed, high wire tension, and medium dielectric flow rate.

43.4.3 Perpendicularity Errors

The perpendicularity of the surfaces at each corner of the test part was checked. For prismatic components, a perpendicularity error of the surfaces is another important dimensional accuracy characteristic, which is also geometric in nature.

The perpendicularity error results given in Table 43.6 illustrate that the surfaces produced by WEDM have about five times higher perpendicularity errors compared to surfaces produced by CNC end milling. The wire lag phenomenon is believed to be responsible for increasing the perpendicularity error of all corners.

The Pareto ANOVA shown in Table 43.7 illustrates that the most significant impendent parameter affecting the flatness error was wire tension (E) ($P = 27.94\%$), followed by pulse gap frequency (C) ($P = 24.62\%$) and wire speed (D) ($P = 15.51\%$). The total of all interaction effects on flatness errors is relatively higher ($P \cong 27.5\%$) compared to that of dimensional errors ($P \cong 10\%$). Therefore, it will be more difficult to control flatness errors through the individual selection of input parameters.

The response graphs for perpendicularity errors are shown in Fig. 43.1c. Based on the *S*/*N* ratio and Pareto ANOVA, it was found that the combination for achieving a low linear dimensional error value was $A_2B_1C_1D_0E_1F_0$; that is, a high discharge current, medium pulse rate, medium pulse gap frequency, low wire speed, medium wire tension, and low dielectric flow rate.

A summary of the optimum levels for the WEDM input parameters is given in Table 43.8. It is clear from the information shown in Table 43.8 that different input parameters are required to be kept at different levels in order to optimize each dimensional accuracy characteristic. This emphasizes the problem of optimizing all three dimensional accuracy characteristics all at once.

Table 43.7 Pareto ANOVA for perpendicularity error

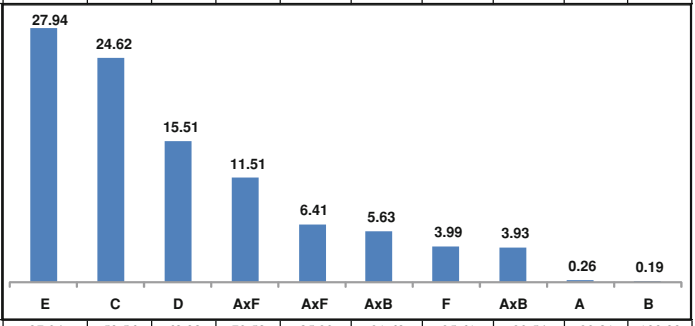
Sum at factor level	Factor and interaction									
	A	B	AxB	AxB	F	AxF	AxF	C	D	E
0	103.15	102.67	138.68	122.96	133.39	80.77	75.78	64.23	56.78	72.496
1	107.31	111.93	99.33	82.47	102.62	152.38	119.31	171.84	135.67	173.031
2	114.69	110.55	87.14	119.72	89.14	92.00	130.06	89.09	132.71	65.986
Sum of sq. of difference (S)	204.83	150.03	4353.34	3037.98	3086.27	8900.77	4957.38	19044.37	11996.45	21608.15
Contribution ratio (%)	0.26	0.19	5.63	3.93	3.99	11.51	6.41	24.62	15.51	27.94
										
	E	C	D	AxF	AxF	AxB	F	AxB	A	B
Cumulative contribution	27.94	52.56	68.08	79.58	85.99	91.62	95.61	99.54	99.81	100.00
Check on significant interaction			AxF two-way table							
Optimum combination of significant factor level			A2 B1 C1 D0 E1 F0							

Table 43.8 Summary of optimum levels for input variables

Dimensional accuracy Characteristics	Optimum levels					
	A	B	C	D	E	F
Linear dimensional error	1	0	2	2	0	2
Flatness error	0	2	0	0	2	1
Perpendicularity error	2	1	1	0	1	0

43.5 Concluding Remarks

From the experimental work conducted and the subsequent analysis, the following conclusions can be drawn:

- The dimensional accuracy achievable in WEDM is not as high as anticipated and its precision level is far less than CNC end milling.
- Of the six input parameters considered, wire tension showed the greatest overall affect on three dimensional accuracy characteristics, therefore, its value should be chosen carefully.
- The problem of erosion of the corner shapes caused by the wire lag phenomenon remains; consequently requires more research and their practical applications.
- Different input parameters are required to be kept at different levels for optimizing each dimensional accuracy characteristic, which highlights the problem of simultaneously optimizing a number of dimensional accuracy characteristics. A hybrid model can be developed to tackle this problem.

References

1. Black JT, Koher RA (2008) Materials and processes in manufacturing, 10th edn. Wiley, Hoboken
2. Hocheng H, Lei WT, Hsu HS (1997) Preliminary study of material removal in electrical discharge machining of SiC/Al. J Mater Process Technol 63:813–818
3. Lee SH, Li X (2001) Study of the effect of machining parameters on the machining characteristics in electrical discharge machining of Tungsten carbide. J Mater Process Technol 115:334–358
4. Yan M, Lai Y (2007) Surface quality improvement of wire-EDM using a fine-finish power supply. Int J Mach Tools Manuf 47:1686–1694
5. Williams RE, Rajurkar KP (1991) Study of wire electrical discharge machining surface characteristics. J Mater Process Technol 28:486–493
6. Tasi K, Wang P (2001) Comparison of neural network models on material removal rate in electrical discharge machining. J Mater Process Technol 117:111–124
7. Dauw DF, Beltrami ETHI (1994) High-precision wire-EDM by online positioning control. Ann CIRP 43(1):193–197
8. Yan M, Huang P (2004) Accuracy improvement of wire-EDM by real-time wire tension control. Int J Mach Tools Manuf 44:807–814
9. Han F, Zhang J, Soichiro I (2007) Corner error simulation of rough cutting in wire EDM. Precis Eng 31:331–336
10. Islam MN, Rafai NH, Subramanian SS (2010) An investigation into dimensional accuracy achievable in wire-cut electrical discharge machining. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, pp 2476–2481
11. Ross PJ (1988) Taguchi techniques for quality engineering. McGraw-Hill, New York
12. Park SH (1996) Robust design and analysis for quality engineering. Chapman & Hall, London
13. Groover MP (2010) Fundamentals of modern manufacturing: materials, processes, and systems, 4th edn. Wiley, Danvers
14. Gladman CA (1972) Geometric analysis of engineering designs, 2nd edn. Australian Trade Publishing Pty. Ltd., Sydney
15. Bjørke Ø (1989) Computer-aided tolerancing, 2nd edn. ASME Press, New York
16. Farmer LE (1999) Dimensioning and tolerancing for function and economic manufacture. Blueprint Publications, Sydney
17. ASME Y (2009) Dimensioning and tolerancing. ASME, New York
18. Islam MN (1995) A CMM-based geometric accuracy study of CNC end milling operations. In: Proceedings of 6th International Conference on Manufacturing Engineering, Melbourne, 29 Nov–1 Dec 1995, pp 835–841

Chapter 44
Nash Game-Theoretic Model for Optimizing Pricing and Inventory Policies in a Three-Level Supply Chain

Yun Huang and George Q. Huang

Abstract This paper aims to coordinate pricing and inventory decisions in a multi-level supply chain composed of multiple suppliers, one manufacturer and multiple retailers. We model this problem as a three-level nested Nash game where all the suppliers formulate the bottom-level Nash game, the whole supplier sector play the middle-level Nash game with the manufacturer, and both sectors as a group player formulate the top-level Nash game with the retailers. Analytical method and solution algorithm are developed to determine the equilibrium of the game. A numerical study is conducted to understand the influence of different parameters on the decisions and profits of the supply chain and its constituent members. Several interesting research findings have been obtained.

44.1 Introduction

In the decentralized supply chain, inconsistence existed between local objectives and the total system objectives make the supply chain lose its competitiveness increasingly [10]. Many researchers have recommended organizational coordination for managing supply chain efficiently [1, 3, 5, 11]. Integrating pricing with inventory decisions is an important aspect to manufacturing and retail industries.

Coordinating pricing and inventory decisions of supply chain (CPISC) has been studied by researchers at about 50 years ago. Weng and Wong [14] and Weng [13]

Y. Huang (✉) · G. Q. Huang
Department of Industrial and Manufacturing Systems Engineering, University of Hong Kong, Pokfulam, Hong Kong
e-mail: huangyun@hku.hk

G. Q. Huang
e-mail: gqhuang@hku.hk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_44,
© Springer Science+Business Media B.V. 2011

propose a model of supplier-retailer relationship and confirm that coordinated decisions on pricing and inventory benefit both the individual chain members and the entire system. Reference [2] analyzes the problem of coordinating pricing and inventory replenishment policies in a supply chain consisting of a wholesaler, one or more geographically dispersed retailers. They show that optimally coordinated policy could be implemented cooperatively by an inventory-consignment agreement. Prafulla et al. [7] present a set of models of coordination for pricing and order quantity decisions in a one manufacturer and one retailer supply chain. They also discuss the advantages and disadvantages of various coordination possibilities. The research we have discussed above, mainly focus on coordination of individual entities or two-stage channels. In reality, a supply chain usually consists of multiple firms (suppliers, manufacturers, retailers, etc.). Jabber and Goyal [6] consider coordination of order quantity in a multiple suppliers, a single vendor and multiple buyers supply chain. Their study focuses on coordination of inventory policies in three-level supply chains with multiple firms at each stage.

Recently, Game Theory has been used as an alternative to analyze the marketing and inventory policies in supply chain wide. Weng [12] study a supply chain with one manufacturer and multiple identical retailers. He shows that the Stackelberg game guaranteed perfect coordination considering quantity discounts and franchise fees. Yu et al. [15] simultaneously consider pricing and order intervals as decisions variable using Stackelberg game in a supply chain with one manufacturer and multiple retailers. Esmaeili et al. [4] propose several game models of seller-buyer relationship to optimize pricing and lot sizing decisions. Game-theoretic approaches are employed to coordinate pricing and inventory policies in the above research, but the authors still focus on the two-stage supply chain.

In this paper, we investigate the CPISC problem in a multi-level supply chain consisting of multiple suppliers, a single manufacturer and multiple retailers. The manufacturer purchases different types of raw materials from his suppliers. Single sourcing strategy is adopted between the manufacturer and the suppliers. Then the manufacturer uses the raw materials to produce different products for different independent retailers with limited production capacity. In this supply chain, all the chain members are rational and determine their pricing and replenishment decisions to maximize their own profits non-cooperatively.

We describe the CPISC problem as a three-level nested Nash game with respect to the overall supply chain. The suppliers formulate the bottom-level Nash game and as a whole play the middle-level Nash game with the manufacturer. Last, the suppliers and the manufacturer being a group formulate the top-level Nash game with the retailers. The three-level nested Nash game settles an equilibrium solution such that any chain member cannot improve his profits by acting unilaterally without degrading the performance of other players. We propose both analytical and computational methods to solve this nested Nash game.

This paper is organized as follows. The next section gives the CPISC problem description and notations to be used. Section 44.3 develops the three-level nested Nash game model for the CPISC problem. Section 44.4 proposes the analytical and computational methods used to solve the CPISC problem in Sect. 44.3.

In Sect. 44.5, a numerical study and corresponding sensitivity analysis for some selected parameters have been presented. Finally, this paper concludes in Sect. 44.6 with some suggestions for further work.

44.2 Problem Statement and Notations

44.2.1 Problem Description and Assumptions

In the three-level supply chain, we consider the retailers facing the customer demands of different products, which can be produced by the manufacturer with different raw materials purchased from the suppliers. These non-cooperative suppliers reach an equilibrium and as a whole negotiate with the manufacturer on their pricing and inventory decisions to maximize their own profits. After the suppliers and the manufacturer reach an agreement, the manufacturer will purchase these raw materials to produce different products for the retailers. Negotiation will also be conducted between the manufacturer and the retailers on their pricing and inventory decisions. When an agreement is achieved between them, the retailers will purchase these products and then distribute to their customers. We then give the following assumptions of this paper:

1. Each retailer only sells one type of product. The retailers' markets are assumed to be independent of each other. The annual demand function for each retailer is the decreasing and convex function with respect to his own retail price.
2. Solo sourcing strategy is adopted between supplies and manufacturer. That is to say, each supplier provides one type of raw materials to the manufacturer and the manufacturer purchases one type of raw material from only one supplier.
3. The integer multipliers mechanism [9] for replenishment is adopted. That is, each supplier's cycle time is an integer multiplier of the cycle time of the manufacturer and the manufacturer's replenishment time is the integer multipliers of all the retailers.
4. The inventory of the raw materials for the manufacturer only occurs when production is set up.
5. Shortage are not permitted, hence the annual production capacity is greater than or equal to the total annual market demand [4].

44.2.2 Notations

All the input parameters and variables used in our models will be stated as follow. Assume the following relevant parameters for the retailer:

L Total number of retailers

r_l Index of retailer l

A_{r_l} A constant in the demand function of retailer l, which represents his market scale
e_{r_l} Coefficient of the product's demand elasticity for retailer l
p_{r_l} Retail price charged to the customer by retailer l
D_{r_l} Retailer l's annual demand
R_{r_l} Retailer l's annual fixed costs for the facilities and organization to carry this product
X_{r_l} Decision vectors set of retailer l. $x_{r_l} \in X_{r_l}$ is his decision vector
Z_{r_l} Objective (payoff) function of retailer l

The retailer's decision variables are:
G_{r_l} Retailer l's profit margin
k_{r_l} The integer divisor used to determine the replenishment cycle of retailer l

The manufacturer's relevant parameters are
m Index of manufacturer
h_{mp_l} Holding costs per unit of product l inventory
$h_{mr_{s_v}}$ Holding costs per unit of raw material purchased from supplier v
S_m Setup cost per production
O_m Ordering processing cost per order of raw materials
P_l Annual production capacity product l, which is a known constant
R_m Manufacturer's annual fixed costs for the facilities and organization for the production of this product
p_{m_l} Wholesale price charged by the manufacturer to the retailer l
c_{m_l} Production cost per unit product l for the manufacturer
X_m Decision vectors set of the manufacturer
$x_m \in X_m$ is his decision vector
Z_m Objective (payoff) function of the manufacturer

The manufacturer's decision variables are:
G_{m_l} Manufacturer's profit margin for product l
T Manufacturer's setup time interval

The relevant parameters for the supplier are
V Total number of suppliers
s_v Index of supplier v, $v = 1, 2, \ldots, V$
h_{s_v} Holding costs per unit of raw material inventory for supplier v
c_{s_v} Raw material cost paid by supplier v
R_{s_v} Supplier v's annual fixed costs for the facilities and organization to carry the raw material
O_{s_v} Order processing cost for supplier v per order
δ_{s_v} Usage of supplier v's raw material to produce a unit product l

p_{s_v} Raw material price charged by supplier v to the manufacturer
X_{s_v} Decision vectors set of supplier v. $x_{s_v} \in X_{s_v}$ is his decision vector
Z_{s_v} Objective (payoff) function of supplier v

The supplier's decision variables are
G_{s_v} Supplier v's profit margin
K_{s_v} The integer multiplier used to determine the replenishment cycle of supplier v

44.3 Model Formulation

44.3.1 The Retailers' Model

We first consider the objective (payoff) function Z_{r_l} for the retailers. The retailer's objective is to maximize his net profit by optimizing his profit margin G_{r_l} and replenishment decision k_{r_l}.

As indicated in the fourth point of the assumption in Sect. 44.2.1, the integer multipliers mechanism is employed between the manufacturer and the retailers. Since the setup time interval for the manufacturer is assumed to be T, the replenishment cycle for retailer l is $T/k_{r_l}.k_{r_i}$ should be a positive integer. Thus, the annual holding cost is $h_{r_l}TD_{r_l}/2k_{r_l}$ and the ordering process cost is $O_{r_l}k_{r_l}/T$.

The retailer l faces the holding cost, the ordering cost and an annual fixed cost. Therefore, the retailer l's objective function is given by the following equation:

$$\max_{k_{r_l}, G_{r_l}} Z_{r_l} = G_{r_l}D_{r_l} - \frac{TD_{r_l}}{2k_{r_l}}h_{r_l} - \frac{O_{r_l}k_{r_l}}{T} - R_{r_l}, \tag{44.1}$$

Subject to

$$k_{r_l} \in \{1, 2, 3, \ldots\}, \tag{44.2}$$

$$G_{r_l} = p_{r_l} - p_{m_l}, \tag{44.3}$$

$$D_{r_l} = A_{r_l} - e_{r_l}p_{r_l}, \tag{44.4}$$

$$G_{r_l} \geq 0, \tag{44.5}$$

$$0 \leq D_{r_l} \leq P_l. \tag{44.6}$$

Constraint (44.2) shows the demand function. Constraint (44.3) gives the value of the divisor used to determine the retailer l's replenishment cycle time. Constraint (44.4) indicates the relationship between the prices (the retail price and the wholesale price) and retailer l's profit margin. Constraint (44.5) ensures that the value of G_{r_l} is nonnegative. Constraint (44.6) gives the bounds of the annual demand, which cannot exceed the annual production capacity of the product.

44.3.2 The Manufacturer's Model

The manufacturer's objective is to determine his decision vector x_m, composed of the profit margins for all the products G_{m_l} and the setup time interval for production T, to maximize his net profit.

The manufacturer faces annual holding costs, setup and ordering costs, and an annual fixed cost. The annual holding cost for the manufacturer is composed of two parts: the cost of holding raw materials used to convert to products, the cost of holding products. During the production portion, the average inventory of raw material v used for product l is $\delta_{s_{lv}} TD_{r_l}/2$. The production time in a year is D_{r_l}/P_l. During the non-production portion of the cycle, the raw materials inventory drops to zero and the holding cost is zero according to our assumption. Hence, the annual holding cost for raw material v is $h_{mr_{s_v}} \delta_{s_{lv}} T(D_{r_l})^2/2P_l$. The annual inventory for product l's is given by $\frac{T}{2} D_{r_l} \left(1 + \frac{1}{k_{r_l}} - \frac{D_{r_l}}{P_l}\right)$ (as suggested by [9]). The setup cost S_m and ordering cost O_m occur at the beginning of each production. Thus, we can easily derive the manufacturer's objective (payoff) function Z_m:

$$\max_{G_{m_1},\ldots,G_{m_l},\ldots,G_{m_L},T} Z_m = \sum_l G_{m_l} D_{r_l} - \frac{T}{2} \sum_l \left(D_{r_l} \left(1 + \frac{1}{k_{r_l}} - \frac{D_{r_l}}{P_l}\right) h_{mp_l} \right) - \sum_l \sum_v \frac{\delta_{s_{lv}} T (D_{r_l})^2}{2P_l} h_{mr_{s_v}} - \frac{S_m + O_m}{T} - R_m \tag{44.7}$$

Subject to

$$G_{m_l} = p_{m_l} - \sum_v \delta_{s_{lv}} p_{s_v} - c_{m_l}, \quad \text{for each } l = 1, 2, \ldots, L \tag{44.8}$$

$$G_{m_l} \geq 0 \quad \text{for each } l = 1, 2, \ldots, L \tag{44.9}$$

$$T > 0. \tag{44.10}$$

Constraint (44.8) gives the relationship between the price (the wholesale price and the raw material price) and the manufacturer's profit margin. Constraint (44.9) and (44.10) ensure that the values of G_{m_l} and T are nonnegative.

44.3.3 The Suppliers' Model

Each supplier's problem is to determine an optimal decision vector x_{s_v} (including replenishment decisions K_{s_v} and profit margin G_{s_v}) to maximize his net profit.

The integer multipliers mechanism is adopted between the suppliers and the manufacturer. So the replenishment cycle time for supplier v is $K_{s_v} T$. The raw material inventory drops every T year by $T \sum_l \delta_{s_{lv}} D_{r_l}$ starting from $(K_{s_v} - 1)T \sum_l \delta_{s_{lv}} D_{r_l}$. Therefore, the holding cost is

$$(K_{s_v}-1)T\sum_l \delta_{s_{lv}} D_{r_l}(Th_{s_v}) + (K_{s_v}-2)T\sum_l \delta_{s_{lv}} D_{r_l}(Th_{s_v}) + \cdots + T\sum_l \delta_{s_{lv}} D_{r_l}(Th_{s_v}),$$

which is equal to $\frac{(K_{s_v}-1)T\sum_l \delta_{s_{lv}} D_{r_l}}{2} h_{s_v}$. The supplier faces holding costs, ordering costs, and an annual fixed cost. Thus, the supplier v's objective (payoff) function Z_{s_v} is

$$\max_{K_{s_v}, G_{s_v}} Z_{s_v} = G_{s_v}\sum_l \delta_{s_{lv}} D_{r_l} - \frac{(K_{s_v}-1)T\sum_l \delta_{s_{lv}} D_{r_l}}{2} h_{s_v} - \frac{O_{s_v}}{K_{s_v}T} - R_{s_v} \tag{44.11}$$

Subject to

$$K_{S_V} \in \{1, 2, 3, \ldots\}, \tag{44.12}$$

$$G_{s_v} = p_{s_v} - c_{s_v}, \tag{44.13}$$

$$G_{s_v} \geq 0. \tag{44.14}$$

Constraint (44.12) gives the value of supplier's multiplier used to determine his replenishment cycle time. Constraint (44.13) indicates the relationship between the raw material price and the supplier's profit margin. Constraint (44.14) ensures the non-negativeness of G_{s_v}.

44.4 Solution Algorithm

In this paper, we mainly based on analytical theory used by [6] to compute Nash equilibrium. In order to determine the three-level nested Nash equilibrium, we first use analytic method to calculate the best reaction functions of each player and employ algorithm procedure to build the Nash equilibrium.

44.4.1 Reaction Functions

We express the retailer l's demand function by the corresponding profit margins. Substituting (44.3), (44.8), (44.13), we can rewrite (44.5) as

$$D_{r_l} = A_{r_l} - e_{r_l}\left(G_{r_l} + G_{m_l} + \sum_v \delta_{s_{lv}}(G_{s_v} + c_{s_v}) + c_{m_l}\right). \tag{44.15}$$

Now suppose that the decision variables for suppliers and manufacturer are fixed.

For retailer l, the best reaction k_{r_l} can be expressed as Viswanathan and Wang [12]:

$$k_{r_l}^* = \left\lfloor \left(1 + \sqrt{1 + \frac{2T^2 h_{r_l} D_{r_l}}{O_{r_l}}}\right)/2 \right\rfloor. \tag{44.16}$$

Here, we define $\lfloor a \rfloor$ as the largest integer no larger than a.

Set the first derivative of Z_{r_l} with respect to G_{r_l} equal to zero. Then G_{r_l} can be obtained as

$$G_{r_l}^* = \frac{C_l}{2e_{r_l}} + \frac{h_{r_l} T}{4k_{r_l}}, \tag{44.17}$$

where $C_l = A_{r_l} - e_{r_l}\left(G_{m_l} + \sum_v \delta_{s_{lv}}(G_{s_v} + c_{s_v}) + c_{m_l}\right)$.

Assume that the decision variables for the suppliers and the retailers are fixed. The manufacturer's problem of finding the optimal setup interval in this case can be derived from $\frac{\partial Z_m}{\partial T} = 0$.

Thus,

$$T^* = \sqrt{\frac{S_m + O_m}{\frac{1}{2}\sum_l \left(D_{r_l}\left(1 + \frac{1}{k_{r_l}} - \frac{D_{r_l}}{P_l}\right)h_{mp_l}\right) + \sum_l \sum_v \frac{\delta_{s_{lv}}\left(D_{r_l}\right)^2}{2P_l} h_{mr_{s_v}}}} \tag{44.18}$$

The optimal G_{m_l} can be obtained from the first order condition of Z_m:

$$G_{m_l}^* = -\frac{W_{m_l} - \frac{h_{mp_l} T}{2}\left(1 + \frac{1}{k_{r_l}}\right)}{2 - \frac{h_{mp_l} T}{2}\left(1 + \frac{1}{k_{r_l}} + \frac{2e_{r_l}}{P_l}\right) + \frac{e_{r_l} T}{P_l}\sum_v \delta_{s_{lv}} h_{mr_{s_v}}} + W_{m_l}, \tag{44.19}$$

where $W_{m_l} = \frac{A_{r_l}}{e_{r_l}} - \left(G_{r_l} + \sum_v \delta_{s_{lv}}(G_{s_v} + c_{s_v}) + c_{m_1}\right)$.

Lastly, we consider the reaction functions for the suppliers. Suppose that the decision variables for retailers and manufacturer are fixed.

The optimal K_{s_v} can be expressed as follows:

$$K_{s_v}^* = \left\lfloor \left(1 + \sqrt{1 + \frac{8O_{s_v}}{T^2 h_{s_v} \sum_l \delta_{s_{lv}} D_{r_l}}}\right)/2 \right\rfloor. \tag{44.20}$$

The necessary condition to maximize the supplier's net profit Z_{s_v} is $\frac{\partial Z_{s_v}}{\partial G_{s_v}} = 0$.

We can obtain

$$G_{s_v} = \frac{-\sum_l \delta_{s_{lv}} e_{r_l} \sum_{\substack{u=1,\ldots,V \\ u \neq v}} \delta_{s_{lu}} G_{s_u} + \sum_l \delta_{s_{lv}} E_l}{2\sum_l \delta_{s_{lv}}^2 e_{r_l}} + \frac{(K_{s_v} - 1)T}{4} h_{s_v}, \tag{44.21}$$

where $E_l = A_{r_l} - e_{r_l}\left(G_{r_l} + G_{m_l} + \sum_v \delta_{s_{lv}} c_{s_v} + c_{m_l}\right)$.

44.4.2 Algorithm

We denote $X_{r_l} = (G_{r_l}, k_{r_l}), X_m = (G_{m_1}, G_{m_2}, \ldots, G_{m_L}, T)$ and $X_{s_v} = (G_{s_v}, K_{s_v})$ as the sets of decision vectors of retailer l, manufacturer and supplier v, respectively. $X_s = X_{s_1} \times \cdots \times X_{s_v}$, $X_r = X_{r_1} \times \cdots \times X_{r_L}$, $X_{ms} = X_s \times X_m$ and $X = X_{ms} \times X_r$ are the strategy profile sets of the suppliers, the retailers, the suppliers and the manufacturer, and all the chain members.

We present the following algorithm for solving the three-level nested Nash game model:

Step 0 Initialize $x^{(0)} = \left(\left(x_s^{(0)}, x_m^{(0)}\right), x_r^{(0)}\right)$ in strategy set.

Step 1 Denote $x_{r_{-l}}^{(0)}$ as the strategy profile of all the chain members in $x^{(0)}$ except for retailer l. For each retailer l, fixed $x_{r_{-l}}^{(0)}$, find out the optimal reaction $x_{r_l}^* = \left(G_{r_l}^*, k_{r_l}^*\right)$ by (44.16) and (44.17) to optimize the retailer l's payoff function Z_{r_l} in its strategy set I_{r_l}.

Step 2 Denote $x_{-m}^{(0)}$ as the strategy profile of all the chain members in $x^{(0)}$ except for the manufacturer. Fixed $x_{-m}^{(0)}$, find out the optimal reaction $x_m^* = \left(G_{m_1}^*, G_{m_2}^*, \ldots, G_{m_L}^*, T^*\right)$ by (44.18) and (44.19), to optimize the profit function Z_m in its strategy set I_m.

Step 3 Denote $x_{s_{-v}}^{(0)}$ as the strategy profile of all the chain members in $x^{(0)}$ except for supplier v. For each supplier v, fixed $x_{s_{-v}}^{(0)}$, find out the optimal reaction $x_{s_v}^* = \left(G_{s_v}^*, K_{s_v}^*\right)$ by (44.20) and (44.21) to optimize the profit function Z_{s_v} in its strategy set I_{s_v}. If $\left\|x_s^* - x_s^{(0)}\right\| = 0$, the bottom level Nash Equilibrium x_s^* obtained, Go step 4. Otherwise, $x_s^{(0)} = x_s^*$, repeat step 3.

Step 4 $x_{ms}^* = \left(x_s^*, x_m^*\right)$. If $\left\|x_{ms}^* - x_{ms}^{(0)}\right\| = 0$, the middle level Nash equilibrium x_{ms}^* obtained, Go step 5. Otherwise, $x_{ms}^{(0)} = x_{ms}^*$, go step 2.

Step 5 $x^* = \left(x_{ms}^*, x_r^*\right)$. If $\left\|x^* - x^{(0)}\right\| = 0$, the above level Nash equilibrium x^* obtained. Output the optimal results and stop. Otherwise, $x^{(0)} = x^*$, go step 1.

44.5 Numerical Example and Sensitive Analysis

In this section, we present a simple numerical example to demonstrate the applicability of the proposed solution procedure to our game model. We consider a supply chain consisting of three suppliers, one single manufacturer and two retailers. The manufacturer procures three kinds of raw materials from the three suppliers. Then the manufacturer uses them to produce two different products and

distributes them to two retailers. The related input parameters for the base example are based on the suggestions from other researchers [8, 10, 14]. For example, the holding cost per unit final product at any retailer should be higher than the manufacturer's. The manufacturer's setup cost should be much larger than any ordering cost. These parameters for the based example are given as: $h_{s_1} = 0.01$, $h_{s_2} = 0.008$, $h_{s_3} = 0.002$, $O_{s_1} = 12$, $O_{s_2} = 23$, $O_{s_2} = 13$, $c_{s_1} = 0.93$, $c_{s_2} = 2$, $c_{s_3} = 6$, $h_{mr_{s_1}} = 0.05$, $h_{mr_{s_2}} = 0.02$, $h_{mr_{s_3}} = 0.04$, $h_{mp_1} = 0.5$, $h_{mp_2} = 1$, $c_{m_1} = 15$, $c_{m_2} = 25$, $\delta_{s_{11}} = 1$, $\delta_{s_{12}} = 2$, $\delta_{s_{13}} = 3$, $\delta_{s_{21}} = 5$, $\delta_{s_{22}} = 4$, $\delta_{s_{23}} = 2$, $S_m = 1,000$, $O_m = 50$, $P_1 = 500,000$, $P_2 = 300,000$, $h_{r_1} = 1$, $h_{r_2} = 2$, $A_{r_1} = 200,000$, $A_{r_2} = 250,000$, $e_{r_1} = 1,600$, $e_{r_2} = 1,400$, $O_{r_1} = 40$, $O_{r_2} = 30$. And the fixed cost for all the players are 1,000.

By applying the above solution procedure in Sect. 5.2, the optimal results for the suppliers, the manufacturer and the retailers are shown in Table 44.1 In order to ensure that our conclusions are not based purely on the chosen numerical values of the base example, we also conduct some sensitive analysis on some parameters, including the market related parameter, the production related parameter and the raw material related parameter. Some meaningful managerial implications can be drawn:

Firstly, the increase of market parameter e_{r_1} will reduce the retailer 1's profit, but benefit the other retailer. When e_{r_1} increases, the change of the retailer 1's demand is more sensitive to the change of his retail price compared with the base example. The retailer 1's profit can be less reduced by lowering down his retail price. But his market demand cannot be increased, which makes the manufacturer

Table 44.1 Results for suppliers, manufacturer and retailers under different parameters

(a) Replenishment decisions

		K_{s_1}	K_{s_2}	K_{s_3}	k_{r_1}	k_{r_2}	T
Base example		1	1	2	6	11	0.2691
e_{r_1}	1,920	1	1	2	6	11	0.2695
S_m	1,500	1	1	1	7	14	0.3252
c_{m_1}	1.395	1	1	2	6	11	0.2707

(b) Pricing decisions

		p_{s_1}	p_{s_2}	p_{s_3}	p_{m_1}	p_{m_2}	p_{r_3}	p_{r_3}
Base example		1.77	6.15	15.15	79.37	101.96	102.20	140.28
e_{r_1}	1,920	1.99	5.99	13.58	71.98	100.42	88.09	139.51
S_m	1,500	1.76	6.20	15.27	79.55	101.61	102.28	140.10
c_{m_1}	1.395	2.23	6.09	15.01	79.44	103.50	102.23	141.05

(c) Product demand and profits

		$D_{r_1}(10^4)$	$D_{r_2}(10^4)$	$Z_{s_1}(10^5)$	$Z_{s_2}(10^5)$	$Z_{s_3}(10^5)$	$Z_m(10^5)$	$Z_{r_3}(10^5)$	$Z_{r_3}(10^5)$
Base example		3.6486	5.3608	2.5335	11.909	19.804	7.7953	8.0236	18.943
e_{r_1}	1,920	3.0870	5.4689	3.2029	11.181	15.305	7.4206	4.7535	19.597
S_m	1,500	3.6345	5.3857	2.5306	12.081	20.084	6.7750	7.8407	18.269
c_{m_1}	1.395	3.6426	5.2531	2.4945	11.568	19.296	7.5513	7.9950	18.149

seek for higher profit from other product to fill up the loss deduced by this product/retail market. It is good news to the other product/retailer, because the manufacturer will lower down his wholesale price to stimulate this market demand.

Secondly, when the manufacturer's setup cost S_m increases, the manufacturer's profit decreases more significantly than the retailers', while some suppliers' profits increase. The increase of S_m makes the manufacturer produce more product with higher profit margin (product 2) and reduce the production of lower profitable product (product 1). The usage of raw materials increases as the change of the manufacturer's production strategy. At the same time, some suppliers bump up their prices, thus bringing higher profits to them.

Thirdly, the impact of the increase of supplier 1's raw material cost c_{s_1} on his own profit may not as significant as that on the other suppliers'. The increase of c_{s_1} makes the supplier 1 raise his raw material price and result in an increase cost in final products, as well as the decrease in market demands. Hence, the other suppliers will reduce their prices to keep the market and optimize their individual profits. Supplier 1 has the much lower profit margin than other suppliers, so he will not reduce his profit margin. Hence, the supplier 1's profit decreases least.

Lastly, when the market parameter e_{r_1}, the manufacturer's setup cost S_m, or the supplier's raw material cost c_{s_1} increase, the manufacturer's setup time interval will be lengthened. A higher e_{r_1} or c_{s_1} results in the total market demands decrease, as well as a lower inventory consumption rate. The increase of S_m makes the manufacturer's cost per production hike up. Hence, the manufacturer has to conduct his production less frequently.

44.6 Conclusion

In this paper, we have considered coordination of pricing and replenishment cycle in a multi-level supply chain composed of multiple suppliers, one single manufacturer and multiple retailers. Sensitive analysis has been conducted on market parameter, production parameter and raw material parameter. The results of the numerical example also show that: (a) when one retailer's market becomes more sensitive to their price, his profit will be decreased, while the other retailer's profit will increase; (b) the increase of the manufacturer's production setup cost will bring losses to himself and the retailers, but may increase the profits of some suppliers; (c) the increase of raw material cost causes losses to all the supply chain members. Surprisingly, the profit of this raw material's supplier may not decrease as significant as the other suppliers'; (d) the setup time interval for the manufacture will be lengthened as the increase of the retailer's price sensitivity, the manufacturer's setup cost or the supplier's raw material cost.

However, this paper has the following limitations, which can be extended in the further research. Although this paper considers multiple products and multiple retailers, the competition among them is not covered. Under this competition, the demand of one product/retailer is not only the function of his own price, but also

the other products'/retailers' prices. Secondly, the suppliers are assumed to be selected and single sourcing strategy is adopted. In fact, either supplier selection or multiple sourcing is an inevitable part of supply chain management. Also, we assume that the production rate is greater than or equal to the demand rate to avoid shortage cost. Without this assumption, the extra cost should be incorporated into the future work.

References

1. Adams D (1995) Category management—a marketing concept for changing times. In: Heilbrunn J (ed) Marketing encyclopedia: issues and trends shaping the future. NTC Business Books, Lincolnwood, IL
2. Boyaci T, Gallego G (2002) Coordinating pricing and inventory replenishment policies for one wholesaler and one or more geographically dispersed retailers. Int J Prod Econ 77(2):95–111
3. Curran TA, Ladd A (2000) SAP/R3 business blueprint: understanding enterprise supply chain management. Prentice Hall, Upper Saddle River
4. Esmaeili M, Aryanezhad MB, Zeephongsekul P (2008) A game theory approach in seller-buyer supply chain. Eur J Oper Res 191(2):442–448
5. Huang Y, Huang GQ (2010) Game-theoretic coordination of marketing and inventory policies in a multi-level supply chain. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July, London, UK, pp 2033–2038
6. Jaber MY, Goyal SK (2008) Coordinating a three-level supply chain with multiple suppliers, a vendor and multiple buyer. Int J Prod Econ 11(6):95–103
7. Joglekar P, Tavana M, Rappaport J (2006) A comprehensive set of models of intra and inter-organizational coordination for marketing and inventory decisions in a supply chain. Int J Integr Supply Manag 2(3):251–284
8. Liu B (1998) Stackelberg-Nash equilibrium for multilevel programming with multiple followers using genetic algorithms. Comput Math Appl 36(7):79–89
9. Lu L (1995) A one-vendor multi-buyer integrated inventory model. Eur J Oper Res 81(2):312–323
10. Porter M (1985) Competitive advantage: creating and sustaining superior performance. The Free Press, New York
11. Sandoe K, Corbitt G, Boykin R (2001) Enterprise integration. Wiley, New York
12. Viswanathan S, Wang Q (2003) Discount pricing decisions in distribution channels with price-sensitive demand. Eur J Oper Res 149:571–587
13. Weng ZK (1995) Channel coordination and quantity discounts. Manag Sci 41(9):1509–1522
14. Weng ZK, Wong RT (1993) General models for the supplier's all-unit quantity discount policy. Naval Res Logist 40(7):971–991
15. Yugang Yu, Liang L, Huang GQ (2006) Leader-follower game in vendor-managed inventory system with limited production capacity considering wholesale and retail prices. Int J Logist Res Appl 9(4):335–350

Chapter 45
Operating Schedule: Take into Account Unexpected Events in Case of a Disaster

Issam Nouaouri, Jean Christophe Nicolas and Daniel Jolly

Abstract In case of a disaster thousands of people may be affected. The needs for medical and surgical treatments overwhelm hospitals' capabilities. A disaster is characterized by different disruptions which perturb largely the execution of the established plans. In hospital and more precisely in operating theatres, the decision-makers have to manage these disruptions in real time. In this setting, we propose a reactive approach in order to optimize the operating rooms scheduling taking into account unexpected events. In this chapter we focus on the insertion of unexpected new victim in the pre-established operating schedule and the overflow of surgical processing time. The purpose is to treat all disruptions and so to save the maximum of human lives. We propose heuristic approach performed by the Cplex solver. Empirical study shows that a substantial aid is obtained by using the proposed approach in case of disaster.

45.1 Introduction

Disasters, such as terrorist attacks, earthquakes, hurricanes and pandemics, often cause a high degree of damage. According to the International Federation of Red Cross and Red Crescent Societies, a disaster is defined as an

I. Nouaouri (✉) · J. C. Nicolas · D. Jolly
Faculté des Sciences Appliquées Technoparc Futura, Univ Lille Nord de France
59000 Lille, France UArtois, LGI2A, 62400 Béthune, France
e-mail: issam.nouaouri@fsa.univ-artois.fr

J. C. Nicolas
e-mail: jchristophe.nicolas@univ-artois.fr

D. Jolly
e-mail: daniel.jolly@univ-artois.fr

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_45,
© Springer Science+Business Media B.V. 2011

exceptional event which suddenly kills or injures a large number of people. In this context, many countries impose to their hospitals to have a disaster plan. For example in the USA, the Joint Commission on the Accreditation of Healthcare Organizations requires US hospitals to have a disaster management plan (DMP) [1]. In other countries like France and some francophone countries, state requirements or laws impose each hospital to have a disaster plan so called white plan [2]. In case of a disaster, this plan is sets in motion (response phase) [3].

During a disaster, emergency managers do have to find an optimal schedule for assigning resources in order to treat a maximum of victims. In this context, we propose to study the optimal allocation of human and material resources in hospitals. We focus more precisely on critical resources: operating rooms and medical staffs. However, in such situation, disruptions can take place, perturbing so the execution of the established plans [4]. The challenge of the operating schedule is to take into account unexpected events which occur during a disaster. In fact, some unexpected victims, needing urgent operations, can arrive to the admitting hospital at any time (example: AZF disaster in Toulouse, France [5]). Furthermore, state of victims can deteriorate or improve during surgical acts, thus reducing or lengthening the assessed surgical processing time. These unforeseen events disrupt the established scheduling (operating schedule) and need to be considered in a reactive way.

Several works were interested in emergency logistics system, especially after the terrorist attack of September 11, 2001 and the Katrina hurricane in 2005. Many studies analyze what happened during the disaster and the resulting logistics problems [3, 6, 7]. These works are based on statistical studies. In order to resolve these problems, other works propose solutions for logistics organization (human and material resources) based on human experiences [8] or on pragmatic approaches [9]. Most of them treat transportation and distribution issues using linear programming [10]. These last 2 years, some works were interested in hospital critical resources such as operating theatres and surgical staffs [11, 12]. Some studies published in the literature address emergency problems in hospitals in normal working times. Most of them focus on operating theatres [13–16] which are considered as a bottleneck in the hospital system. These last years, some works were interested in disaster situations [10, 11, 17–19]. The optimization of operating theatres has become an important issue. However, all these works do not consider operating schedule in case of a disaster.

In this chapter, we deal with reactive operating schedule in case of a disaster. Our purpose is to take into account unexpected events. To achieve this, we propose an approach based on a several-stage model.

In Sect. 45.2 of this chapter, we present the reactive problem we address in a disaster case. Sections 45.3 and 45.4 detail the proposed approach and the problem modelling. Section 45.5 discusses the obtained numerical results. Finally, Sect. 45.6 concludes the chapter and presents possible extensions of this work.

45.2 Problem Description

In case of a disaster, victims are evacuated to an immediate established pre-hospital triage and dispatching structure which is set up near to the damaged zone. This structure guarantees the first aid emergency cares and routes victims to the available hospitals. The triage allows classifying victims according to the urgency of the medical and/or surgical cares they need.

In this chapter, we consider victims that require surgical cares with predefined processing times (according to pathology) and different ready dates in the operating theatre. Each victim is characterized by an emergency level which is defined by the latest start date of its surgical care. Therefore, the surgical care must be planned before the vital prognosis of the victim is being overtaken [2, 8, 20]. Pre-hospital and triage structure communicates to the admitting hospital, via information system [2, 8, 21]: surgical processing time, ready date and emergency level of each victim before its arrival to the hospital.

In hospital, some human and material resources are available. We consider critical resources: surgical staffs and operating rooms. Surgical staffs, their number, and ready dates in the operating theatre are detailed in a pre-established emergency planning. Each surgical staff is assigned to one operating room. In case of a disaster, all operating rooms are considered to be polyvalent.

Basing on these information, the admitting hospital achieves its predictive program (operating schedule) at $t = t_0$. However, some unexpected events can occur at any moment during the execution of this program. In reaction to a disturbance at $t = t_p$, hospital must be able to respond quickly and efficiently, in order to minimize the involved consequences.

In case of a disaster, date of arrival of victims, announced by the pre-hospital triage and dispatching structure, is widely variable and depends on type, location, transport capacities and site organization [3]. Furthermore, some victims may arrive at hospitals at any time without passing by the pre-hospital, triage and dispatching structure [5]. Moreover, during the execution of a surgical care in the operating room, state of victims can deteriorate, thus lengthening the processing time of its surgical care. In this context, we handle the scheduling operating rooms problem while taking into account: the arrival of new (unexpected) victims requiring surgical cares, and the overflow of assessed surgical processing time.

45.3 Proposed Approach

Before presenting our proposed approach, we will first introduce the following notations:Snumber of operating rooms; Nnumber of victims; Hnumber of surgical staffs; Ttime horizon; d_i processing time of surgical care of victim i; dl_i latest start date of surgical care of victim i; rv_i ready date of victim i; rc_h ready date of surgical staff h with respect to the hospital emergency planning; Mvery big positive number.

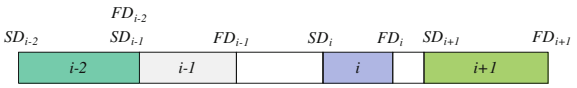


Fig. 45.1 Pre-established schedule

We consider, that the number of rooms is equal to the number of surgical staffs ($H = S$) and each surgical staff is affected to only one operating room. So, we can use h as s and s as h ($h = s$).

45.3.1 Arrival of New Victim

P_0 pre-established operating schedule at $t = t_0$. At $t = t_p$ (date of disruption) a new victim is announced by the pre-hospital, triage and dispatching structure or, arrives directly to the hospital.

In order to minimize disruption effect on P_0, we proceed in several stages. The first one, model (P_1) is stated as follows: Given the operating schedule (P_0), P_1 tries to insert the new victim in an untapped (vacant) range. The model has to satisfy some constraints such as ready dates of surgical staffs and latest start date of the new victim. The surgical processing time of the new victim has to be lower or at more equal to one untapped range. If the new victim cannot be inserted by (P_1), we compute for every operating room a free margin Δg_s from elementary margins Δ_i (Fig. 45.1).

$$\Delta g_s = \sum_{i \in s}^{N} \Delta_i \tag{I}$$

$$\Delta_i = FD\_l_i - FD\_e_i \tag{II}$$

We define: SD_i the start date of surgical care of victim *i*, FD_i the finish date of surgical care of victim *i*.

The latest finish date of surgical care of victim i is given by Eq. III.

$$FD\_l_i = \min(dl_i + d_i, SD_{i+1}) \tag{III}$$

The earliest finish date (date as soon as possible) of surgical care of victim i is given by Eq. IV.

$$FD\_e_i = FD_{i-1} + d_i \tag{IV}$$

If the new victim cannot be inserted by (P_1), the model (P_2) tries to reschedule, from disruption date, the victims (including the new victim) belonging to the operating room which possesses the biggest Δg_s. Current surgical cares will not be interrupted. If the new victim has not been inserted, the model (P_3) tries to reschedule, from disruption date, all surgical cares in all operating rooms. If no solution is found, the new victim cannot be inserted and is proposed for a reorientation to another hospital.

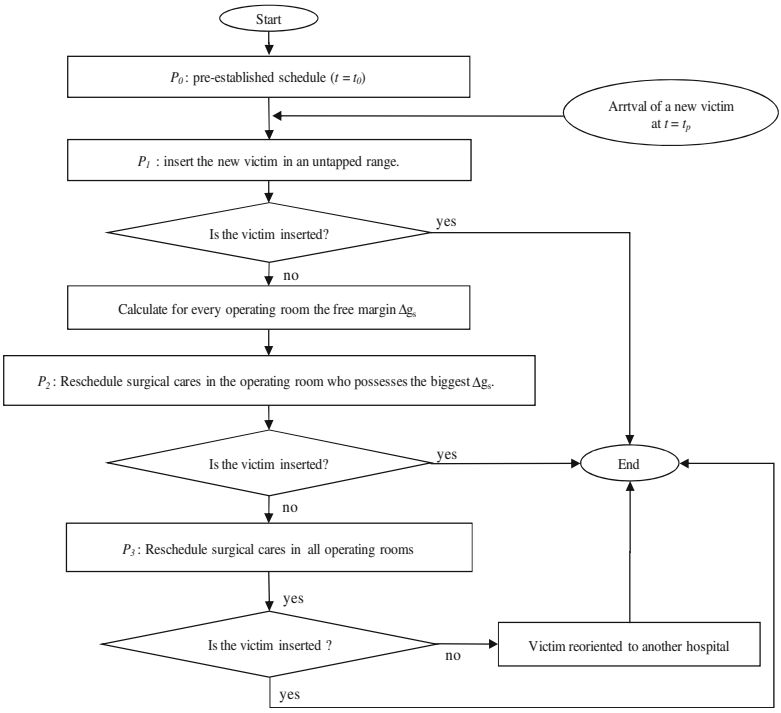

Start
P_0: pre-established schedule ($t = t_0$)
Arrtval of a new victim at $t = t_p$
P_1 : insert the new victim in an untapped range.
Is the victim inserted?
yes
no
Calculate for every operating room the free margin Δg_s
P_2: Reschedule surgical cares in the operating room who possesses the biggest Δg_s.
Is the victim inserted?
yes
End
no
P_3: Reschedule surgical cares in all operating rooms
yes
Is the victim inserted ?
no
Victim reoriented to another hospital
yes

Fig. 45.2 Reactive approach in case of insertion of a new victim

We present in Fig. 45.2, the proposed reactive approach in case of insertion of a new victim in the operating schedule.

45.3.2 Overflow of Assessed Surgical Processing Time

In this case, t_p is the date for which surgical care exceeds the assessed surgical processing time. We proceed in several stages.

In the first stage, we apply "shift right" algorithm (P_4) in order to shift all surgical cares following victim whose state needs more time in operating room, lengthening so its surgical processing time. If (P_4) eliminates other victims from operating schedule, (P_5) tries to reschedule, from disruption date, victims belonging to the same operating room. If disruption cannot be treated, (P_6) reschedule, from disruption date, all surgical cares in all operating rooms. If no solution is found, the least urgent victims, eliminated from operating room, will be oriented to other hospitals. Indeed, these victims have the largest latest start date.

45.4 Problem Modelling

45.4.1 Arrival of New Victim

We propose a three-stage mathematical model. For each stage, an integer linear programming model is developed using ILOG OPL 6.1 Studio.

According to the pre-established schedule (P_0) [4], we define: t_{is} the start date of surgical care of the victim i in the operating room s; $y_{ijs} = 1$ if the surgical care of the victim j follows the surgical care of the victim i in the same operating room s, 0 otherwise; t_p disruption date which is generated in stochastic way; NR number of new victims. In our case, we consider $NR = 1$, because we treat only one unexpected event at a given instant.

Besides, we define the following decision variables: $Z_{kts} = 1$ if the new victim k is assigned to an operating room s at time t, 0 otherwise; st_k start time of surgical care of victim k.

Model 1*: Insertion of the new victim in an untapped range*

In the first stage, we address the optimization problem (P_1).

- The objective function (45.1) maximizes the number of new inserted victims in the operating schedule.

$$Maximize \sum_{k}^{NR} \sum_{t_p}^{T} \sum_{s}^{S} Z_{kts} \tag{45.1}$$

- Constraints (45.2) ensure that each victim is treated only once during the horizon T.

$$\sum_{t_p}^{T} \sum_{s}^{S} Z_{kts} \leq 1 \quad \forall k \in \{1, \ldots, NR\} \tag{45.2}$$

- Constraints (45.3) grantee, for every untapped range, one victim is assigned at most at time t.

$$\sum_{t=t_{is}+d_i/t \geq t_p}^{t_{js}} Z_{kts} \leq y_{ijs} \quad \forall s \in \{1, \ldots, S\} \forall i,\ j \in \{1, \ldots, N\} \forall k \in \{1, \ldots, NR\} \tag{45.3}$$

- Constraints (45.4) impose to satisfy the emergency level of each new victim.

$$st_k - dl_k \sum_{t_p}^{T} \sum_{s}^{S} Z_{kts} - M\left(1 - \sum_{t_p}^{T} \sum_{s}^{S} Z_{kts}\right) \leq 0 \quad \forall k \in \{1, \ldots, NR\} \tag{45.4}$$

- (45.5) Ensures that the duration of surgical care of a new victim is lower or equal to the duration of the untapped range.

$$t_{js}y_{ijs} - (t_{is} + d_i)y_{ijs} \geq d_k \sum_{t_p}^{T} Z_{kts} \quad \forall s \in \{1, \ldots, S\} \forall i,\ j \in \{1, \ldots, N\} \forall k \in \{1, \ldots, NR\} \tag{45.5}$$

- (45.6) and (45.7) verify that surgical care can be realized only when victim and surgical staff are present in the hospital.

$$st_k + M\left(1 - \sum_{t_p}^{T}\sum_{s}^{S} Z_{kts}\right) \geq rv_k \quad \forall k \in \{1, \ldots, NR\} \tag{45.6}$$

$$st_k - rc_s \sum_{t=t_p}^{T} Z_{kts} - M\left(1 - \sum_{t=t_p}^{T} Z_{kts}\right) \geq 0 \quad \forall s \in \{1, \ldots, S\}\ \forall k \in \{1, \ldots, NR\} \tag{45.7}$$

- (45.8) Grantees that surgical care of the new victim cannot be inserted before disruption date t_p.

$$st_k + M\left(1 - \sum_{t_p}^{T}\sum_{s}^{S} Z_{kts}\right) \geq t_p \quad \forall k \in \{1, \ldots, NR\} \tag{45.8}$$

- Constraints (45.9) ensure that every victim k is assigned to an untapped range.

$$(t_{is} + d_i)y_{ijs} \sum_{t_p}^{T} Z_{kts} \leq st_k \leq t_{js}y_{ijs} \sum_{t_p}^{T} Z_{kts} \quad \forall s \in \{1, \ldots, S\}\ \forall k \in \{1, \ldots, NR\} \forall i,\ j \in \{1, \ldots, N\} \tag{45.9}$$

- Constraints (45.10) give the start times of surgical cares.

$$st_k = \sum_{t_p}^{T}\sum_{s}^{S} t.Z_{kts} + \left(1 - \sum_{t_p}^{T}\sum_{s}^{S} Z_{kts}\right)M \quad \forall k \in \{1, \ldots, NR\} \tag{45.10}$$

- Constraints (45.11) ensure the integrality of the variables.

$$Z_{kts} = \{0, 1\} \quad \forall k \in \{1, \ldots, NR\}\ \forall t \in \{t_p, \ldots, T\}\ \forall s \in \{1, \ldots, S\} \tag{45.11}$$

If the new victim has not been inserted, we try to treat it with the model 2 (P_2).

Model 2: Reschedule surgical cares in the operating room which possesses the biggest free margin Δg_s

Before solving this problem, we have to compute: (1) the free margin Δg_s of each operating room s, and (2) the date (As) from which operating room s (surgeon h) is available after t_p. The room s as well as the surgeon h are known beforehand.

The mathematical formulation of this combinatorial optimization problem (P_2) is given by the following linear integer program.

W set of waiting victims (including the new victim) for surgical cares.

The decision variables are defined as follow: $X_{its} = 1$ if the victim i is assigned to an operating room s at time t, 0 otherwise; st_i start time of surgical care of victim i.

- The objective function (45.12) maximizes the number of treated victims in the operating room s after the disruption date t_p.

$$Maximize \sum_{i}^{W} \sum_{t=t_p}^{T} X_{its} \tag{45.12}$$

- Constraints (45.13) ensure that each victim is treated only once during the horizon T.

$$\sum_{t=t_p}^{T} X_{its} \leq 1 \quad \forall i \in W \tag{45.13}$$

- Constraints (45.14) grantee that one victim at most is assigned at time t in the operating room s.

$$\sum_{i}^{W} X_{its} \leq 1 \quad \forall t \in \{t_p, \ldots, T\} \tag{45.14}$$

- Constraints (45.15) impose to satisfy the emergency level of each victim.

$$st_i - dl_i \sum_{t=t_p}^{T} X_{its} - M\left(1 - \sum_{t=t_p}^{T} X_{its}\right) \leq 0 \quad \forall i \in W \tag{45.15}$$

- (45.16) and (45.17) verify that surgical care are realized only when victim and surgical staff are present in the hospital.

$$st_i + M\left(1 - \sum_{t_p}^{T} X_{its}\right) \geq rv_i \quad \forall i \in W \tag{45.16}$$

$$st_i - rc_s \sum_{t=t_p}^{T} X_{its} - M(1 - \sum_{t=t_p}^{T} X_{its}) \geq 0 \quad \forall i \in W \tag{45.17}$$

- Constraints (45.18) verify the availability of the operating room s after the disruption date.

$$st_i \geq A_s.X_{its} - M\left(1 - \sum_{t_p}^{T} X_{its}\right) \quad \forall i \in W \tag{45.18}$$

- Constraints (45.19a), (45.19b) and (45.20) are disjunctive precedence constraints.

$$\sum_{j \neq i}^{W} y_{ijs} \leq 1 \quad \forall i \in W \tag{45.19a}$$

$$\sum_{j \neq i}^{W} y_{jis} \leq 1 \quad \forall i \in W \tag{45.19b}$$

$$\sum_{i}^{W} \sum_{j \neq i}^{W} y_{jis} = \sum_{i}^{W} \sum_{t=t_p}^{T} X_{its} - 1 \tag{45.20}$$

- Equation 45.21 gives the start times of surgical cares.

$$st_i = \sum_{t_p}^{T} t.X_{its} + \left(1 - \sum_{t_p}^{T} t.X_{its}\right) M \quad \forall i \in W \tag{45.21}$$

- Constraints (45.22) impose no overlapping between two successive cares made in same operating room.

$$st_j \geq st_i + y_{ijs} d_i - M(1 - y_{ijs}) \quad \forall i, j \in W \tag{45.22}$$

- Constraints (45.23) and (45.24) ensure the integrality of the variables.

$$X_{its} = \{0, 1\} \quad \forall i \in W \; \forall t \in \{t_p, \ldots, T\} \tag{45.23}$$

$$y_{ijs} = \{0, 1\} \quad \forall i, j \in W \tag{45.24}$$

If the new victim has not been inserted, we try to treat it with the model 3 (P_3).

Model 3: Reschedule surgical cares in all operating rooms

The objective function (45.25) maximizes the number of treated victims in all operating rooms.

$$Maximize \sum_{i}^{W} \sum_{t=t_p}^{T} \sum_{s}^{S} X_{its} \tag{45.25}$$

Under the same kind of constraints used in model 2 but taking into account all operating rooms.

For example constraints (45.2) and (45.3) become (45.26) and (45.27):

$$\sum_{t=tp}^{T} \sum_{s}^{S} X_{its} \leq 1 \quad \forall i \in W \tag{45.26}$$

$$\sum_{i}^{W} X_{its} \leq 1 \quad \forall t \in \{t_p, \ldots, T\} \; \forall s \in \{1, \ldots, S\} \tag{45.27}$$

Numerical results will be presented in Sect. 45.5.

45.4.2 Overflow of Assessed Surgical Processing Time

In order to minimize disruption effect on P_0, we start with applying "shift right" algorithm (P_4). This program is activated after the delay *Re* has been estimated during the surgical process of current surgical care.

Shift right algorithm

1/ Initialization:
s : operating room which present overflow of assessed current surgical processing time.
k : surgical care exceeds the assessed surgical processing time.
I_s : Number of waiting victims for surgical cares in the operating room s.

2/ Shift right all surgical cares following victim whose state needs more time in the operating room, lengthening so its surgical processing time. The shift is equal to the recorded delay R :
Re = concrete spending time of surgical care k – assessed time (d_k).
For ($i = k+1$; $i < I_s$; i++)
$st_i = st_i + Re$
if st_i dl_i
Then
Return to initial operating schedule (P_0) and go -to 3/

3/ *End*

If (P_4) excludes other victims from operating schedule, we go to the model 5 (P_5): *Reschedule surgical cares in the operating room s which present an overflow of assessed surgical processing time*. If (P_5) eliminates other victims from operating schedule, we go to the model 6 (P_6): *Reschedule surgical cares in all operating rooms*.

If no solution is found, the least urgent victims, eliminated from operating room, will be oriented to other hospitals.

45.5 Computational Experiments

In this section, we present the realized computational experiments using the Cplex solver. We run programs on a Cluster consist of 6 workstations Bixeon® of 3.00 GHz processor and 2–4 Go RAM. We evaluate the performances of the proposed reactive approach with different scenarios (generated in stochastic way) described on the following.

45.5.1 Problem Tests

Different disaster situations are considered by varying the number of victims ($N = 25$, 50 and 70) and the duration of surgical cares (given between 30 min and

Table 45.1 Ready dates of surgical staffs according to hospital emergency planning

R	$rc_1^{(mn)}$	$rc_2^{(mn)}$	$rc_3^{(mn)}$	$rc_4^{(mn)}$	$rc_5^{(mn)}$	$rc_6^{(mn)}$	$rc_7^{(mn)}$	$rc_8^{(mn)}$	$rc_9^{(mn)}$	$rc_{10}^{(mn)}$
R_1	0	0	0	30	30	30	60	60	120	120
R_2	0	0	0	30	30	30	60	60	60	120
R_3	0	0	0	30	30	60	60	60	120	120
R_4	0	0	30	30	30	60	60	60	120	120
R_5	0	0	30	30	60	60	60	120	120	120

2 h). Moreover, 10 surgical staffs are available with different ready dates ($R = (rc_1, \ldots, rc_H)$, $H = 10$) according to the hospital emergency planning (Table 45.1). The instance label *PN.S.R* means the problem *P* involves *N* victims, *S* operating rooms and ready dates *R* of surgical staffs.

For example $P70.10.R_3$ denotes a problem of 70 victims and 10 surgical staffs (10 operating rooms) which ready dates in minutes (mn) are given by R_3, thus $rc_1 = 0$, $rc_2 = 0$, $rc_3 = 0$, $rc_4 = 30$, $rc_5 = 30$, $rc_6 = 60$, $rc_7 = 60$, $rc_8 = 60$, $rc_9 = 120$, $rc_{10} = 120$. In this chapter we consider the following instances: $P25.4.R_1$, $P25.6.R_3$, $P25.6.R_5$, $P50.4.R_1$, $P50.6.R_3$, $P50.8.R_3$, $P70.4.R_1$, $P70.6.R_3$, $P70.10.R_3$.

For each instance, we generate a predictive program P_0 (pre-established operating schedule) [4]. We apply for every instance:

- Twenty scenarios by inserting a new victim requiring surgical cares in the operating schedule. One scenario is characterized by the ready date, the processing time and the emergency level of the new victim.
- Twenty scenarios by varying surgical processing time and the victim whose state needs more time in the operating room.

The computational experiments are performed while fixing time horizon $T = \max_{i=1,\ldots,N}(dl_i + d_i)$. Indeed, after this date, no victim can be treated. T is decomposed in elementary periods.

45.5.2 Results

In this section, we propose to detail results given by the insertion of new victims in the operating schedule.

In order to assess the performance of the proposed approach, we calculate for a set of scenarios belonging to one instance, the rate of insertion of the new victims in the operating schedule (V.I (%)).We also compute, the percentage of cases for which disruptions are treated and resolute by the program P_k (V.I.P$_k$ (%)).

$$\text{V.I}(\%) = \frac{\sum_j \text{Inserted new victim } j}{\sum_i \text{New victim } i} \tag{V}$$

$$\text{V.I.P}_k(\%) = \frac{\sum_j \text{New victim } j \text{ inserted by the model } p_k}{\sum_i \text{New victim } i} \tag{VI}$$

Table 45.2 Numerical results in case of arrival of a new victim

Instances	V.I (%)	V.I.P$_1$(%)	V.I.P$_2$(%)	V.I.P$_3$(%)	CPU time		TM(%)
					$T_{\min}$ (s)	$T_{\max}$ (s)	
$P25.4.R_1$	25	5	10	10	26	295	81.25
$P25.6.R_3$	85	10	35	40	47	352	53.17
$P25.6.R_5$	80	10	40	30	43	281	53.33
$P50.4.R_1$	10	0	0	10	524	682	94.82
$P50.6.R_3$	25	0	0	25	609	752	78.73
$P50.8.R_3$	90	15	40	35	43	350	63.79
$P70.4.R_1$	15	0	0	15	869	973	98.30
$P70.6.R_3$	35	0	5	30	826	1239	85.55
$P70.10.R_3$	70	5	25	40	76	1120	59.41

Scenarios are generated in a stochastic way. The results presented in the tables are obtained by solving Model 1 (P_1), Model 2 (P_2) and Model 3 (P_3). We report for each instance the rate of insertion of new victims in the operating schedule (V.I (%)), the percentage of cases for which disruptions are treated and resolved by the program P_k (V.I.P$_k$ (%)), the minimum CPU time ($T_{\min}$) and the maximum CPU time ($T_{\max}$). For each instance, we compute TM (%) the mean occupancy rate of operating rooms for pre-established operating schedule.

$$TM(\%) = \frac{\sum_{i=1}^{N} d_i}{C_{\max}.S} \tag{V1I}$$

$C_{\max}$ is the Makespan given by the pre-established operating schedule.

The proposed approach has allowed inserting new victims in 48% of cases. The solution is obtained between 26 s (minimum) and 21 min (maximum).

Table 45.2 shows that the rate of insertion of new victims varies according to the rate of the mean occupancy rate of operating rooms for pre-established operating schedule (TM (%)) (example: V.I (%) = 25 for TM (%) = 81.25 ($P25.4.R_1$) beside V.I (%) = 85 for TM (%) = 53.17 ($P25.6.R_3$)). In most cases, new victims are inserted by using rescheduling model ((P_2) and (P_3)) (example: in case of $P70.10.R_3$, V.I.P$_2$(%) = 25 and V.I.P$_3$(%) = 40). In fact, reschedule models give good results.

45.6 Conclusion and Future Work

In this chapter, we have addressed a reactive approach in case of a disaster in order to optimize the operating rooms scheduling taking into account unexpected events: (1) insertion of a new victim in the established operating schedule, (2) surgical care exceeds the assessed processing time. Both proposed approaches are based on a three-stage model.

In case of insertion of a new victim in the established operating schedule, the first model tries to insert the new victim in an untapped range. If the victim cannot be treated, the second model tries to reschedule, from disruption date, the victims belonging to the operating room that possesses the biggest free margin. If the new victim has not been inserted, the third model reschedules all surgical cares in all operating rooms. If no solution is founded the victim will be reoriented to another hospital.

In case of one surgical care exceeds the assessed processing time, the first model "shift right" shifts all surgical cares following the concerned victim. If this model excludes other victims from operating schedule, we apply rescheduling models: (1) rescheduling of one operating room and (2) rescheduling of all operating rooms.

The proposed Heuristic approach allows hospital to decide how to take into account disruption, and so to treat a maximum of victims (save the maximum of human life). It is a decision tool used before the arrival of victim (s) at hospital in order to take the best decision quickly and efficiently. Another interesting advantage of this approach is that it tries to resolve the problem in several stages in order to minimize the disruption of pre-established operating schedule.

This approach has been tested on different disaster situations with various scenarios. Further research works should focus on dealing with auxiliary services such as recovery room, post anaesthesia care unit, etc.

Acknowledgments The authors thank the university group and medical partners of the project ARC-IR (France) as well as professionals of the university hospital Charles Nicole (Tunisia) and the General Direction of the military health (Tunisia). We are particularly grateful to Dr. Naoufel Somrani, Dr Henda Chebbi, Pr. Eric Wiel and Dr. Cédric Gozé.

References

1. Lipp M, Paschen H, Daublander M, Bickel-Pettrup R, Dick W (1998) Disaster management in hospitals. Curr Anaesth Crit Care 9:78–85
2. Ministère de la santé et de la solidarité, Plan blanc et gestion de crise. Edition 2006, annexe à la circulaire n°DHOS/CGR/2006/401, France
3. Kimberly A, Cyganik RN (2003) Disaster preparedness in Virginia hospital center—Arlington after Sept 11, 2001. In: Ted Cieslak MD (Section ed) Disaster management and response/cyganik, vol 1. pp 80–86
4. Nouaouri I, Hajri-Gabouj S, Dridi N, Nicolas JCh, Jolly D, Gabouj M (2008) Programmation des interventions de stabilisation dans les salles opératoires: cas d'une catastrophe à effet limité. 7ème Conférence Internationale de MOdélisation et de SIMulation MOSIM'08, Paris, France
5. Ministère de la santé, de la famille et des personnes handicapées, France. Explosion de l'usine AZF de Toulouse le 21 septembre 2001, Rapport de mission, Enseignements et propositions sur l'organisation de soins, 2002, France
6. Kembell-Cook D, Stephenson R (1984) Lessons in logistics from Somalia. Disasters 8:57–66
7. Long DC, Wood DF (1995) The logistics of famine relief. J Bus Logistics 16:213–229
8. Blackwell T, Bosse M Use of innovative design mobile hospital in the medical response to hurricane Katrina, the American college of emergency physicians. doi:10.1016/j.annemergmed.2006.06.037

9. Filoromo C, Macrina D, Pryor E, Terndrup T, McNutt S-D (2003) An innovative approach to training hospital based clinicians for bioterrorist attack. AJIC Practice forum. doi:10.1016/S0196-6553(03)00699-0
10. Sheu J (2007) An emergency logistics distribution approach for quick response to urgent relief demand in disasters. Transp Res Part E 43:687–709
11. Nouaouri I, Nicolas JCh, Jolly D (2009) Scheduling of stabilization surgical cares in case of a disaster. The IEEE International conference on industrial engineering and engineering management (IEEM), Hong Kong, December 2009
12. Welzel, Tyson B, LMD Koenig Kristi, Bey Tareg MD, Visser (2010) Effect of hospital staff surge capacity on preparedness for a conventional mass casualty event. West J Emerg Med 11(2):189–196
13. Hammami S (2004) Aide a la décision dans le pilotage des flux matériels et patients d'un plateau médico-technique. Thèse de doctorat Institut National Polytechnique de Grenoble, France
14. Kuo PC, Schroeder RA, Mahaffey S, Bollinger R (2003) Optimization of operating room allocation using linear programming techniques, the American college of surgeons. Elsevier, USA
15. Lamiri M, Xie X, Dolgui A, Grimaud F (2008) A stochastic model for operating room planning with elective and emergency demand for surgery. Eur J Oper Res 185:1026–1037
16. Roland B, Cordier J-P, Tancrez J-S, Riane F (2009) Recovery beds and blocking in stochastic operating theatres. Conference on industrial engineering and systems management (IESM), Montreal, Canada, May 2009
17. Glaa B (2008) Contribution à la conception et l'optimisation d'un système d'aide à la gestion des urgences. Thèse de doctorat Ecole Centrale de Lille, France
18. Jia H, Ordonez F, Dessouky MM (2007) Solution approaches for facility location of medical supplies for large-scale emergencies. Comput Ind Eng 52:257–276
19. Nouaouri I, Nicolas J-C, Jolly D (2010) Reactive operating schedule in case of a disaster arrival of unexpected victims. Lecture notes in engineering and computer science proceedings of the world congress on engineering, vol 2185(1), pp 2123–2128
20. Dhahri MA (1999) Organisation des secours médicaux dans les situations d'exception en Tunisie. D.E.S.S de médecine d'urgence et de catastrophe, Département d'Anesthésie et de réanimation, Hôpital Militaire de Tunis
21. Hasegawa S, Sato K, Matsunuma S, Miyao M, Okamoto K (2005) Multilingual disaster information system: information delivery using graphic text for mobile phones, AI Society 19(3):265–278

Chapter 46
Dynamic Hoist Scheduling Problem on Real-Life Electroplating Production Line

Krzysztof Kujawski and Jerzy Świątek

Abstract The paper describes a conducted research in order to create the scheduling system for electroplating lines. We are interested in an order driven production organization. This type of production requires scheduling in the real-time. In order to avoid costly scheduling, before the real-time system starts we prepare a set of production scenarios. We use recognition methods to select the most appropriate scenario. Selected scenarios are parameters to a heuristic scheduling algorithm called cyclogram unfolding. We present the short problem notion and algorithm steps. The real-life production line located at Wrocław, Poland is used to explain the algorithm step by step. We analyze the results for historical runs and generated problems. The utilization ratio is introduced in order to judge the quality of created schedules.

46.1 Introduction

In electroplating industry a HSP—the hoist scheduling problem occurs. The problem lies in creating a proper schedule for machines working on a production line. On electroplating lines items are chemically processed. The items are transported from a workstation to a workstation by automated hoists. An electroplating line can produce multiple item types. Each can have its own

K. Kujawski (✉) · J. Świątek
Institute of Informatics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50–370, Wrocław, Poland
e-mail: krzysztof.kujawski@pwr.wroc.pl

J. Świątek
e-mail: Jerzy.swiatek@pwr.wroc.pl

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_46,
© Springer Science+Business Media B.V. 2011

technological process, i.e. sequence of visiting workstations, processing intervals, etc. The production can be organized cyclically or dynamic (the dynamic hoist scheduling problem—DHSP)-order driven. In the cyclic production a static schedule, called a cyclogram, is repeated in order to produce items. Such organization has its advantages, like simplicity and predictability, but lack in flexibility. Having a static schedule requires expensive setups in order to change the produced item type. In DHSP, schedule is created during production. Produced items vary both in type and time of introduction to line. We present the scheduling system, which divides the problem to the *local problem* (Lamothe et al. [9]) where schedules are created with no real-time constraints and the real-time part. Real-time part reacts on new item orders, verifies the feasibility of schedules and implements schedules to line automatons.

In this paper we extend the research from Kujawski [7], connected to real-life electroplating production line located in Wrocław, Poland.

46.2 Dynamic Hoist Scheduling Problem

Electroplating lines cover processed items with thin material coatings using chemical reactions—usually by galvanization. They are automated production systems, which use hoists as transportation. Production line is made of series of workstations. Workstations are capable of performing different elementary chemical reactions or material processing operations. Input product is subjected to several reactions in specific sequence.

Production line consists of: *baths* (*tanks*) and *hoists.* Baths are workstations that perform a certain stage of the chemical or material processing. Groups of uniform workstations are also considered. Some workstations can perform more than one stage in processing. Baths are arranged in line, and they create an axis of hoist movement. Hoists—automatons controlled by schedule. Hoists are capable of transferring products between workstations. Hoists move only in production line axis. Hoists cannot pass each other. A hoist can pick up and put down products to workstation. Hoists cannot pass products between each other. The processing starts when a hoist plunge item to a bath. This happens because of the chemical nature of the processing. When a hoist plunge an item to workstation we refer to it as the operation of putting down of an item. The processing stops, when the hoist picks item out of a bath.

In order to increase flexibility of the production system, some production lines are designed to produce many item types. This is done by composing workstations, which are required to produce the certain type of item. Scheduling item types differ in sequence of visited workstations and times of processing at those workstations. In case of chemical processing, instead of time of processing at a certain workstation, we have quality constrains. It means that an item is immersed in the workstation for at least given minimum time and no longer than given maximum time.

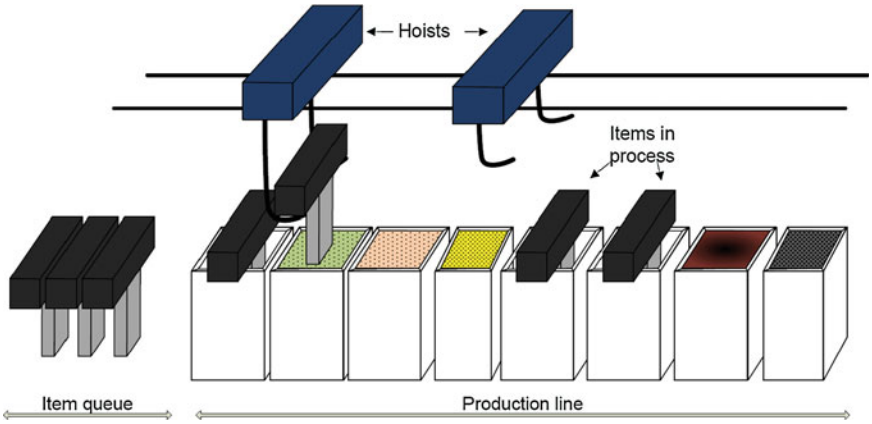


Fig. 46.1 The electroplating line overview

One of the first papers about HSP Phillips and Unger [13] describes the simple application, where the production line has only one hoist and simple sequence of production—always to next workstation. More recent publications Jiyin and Yun [4] Zhou and Liu [15] and Yan [14] expanded the problem to multiple hoists, Leung and Guoqing [10] and Mak et al. [12] an arbitrary production sequence, groups of workstations Frohlich and Steneberg S C [1] and multifunctional baths Mak et al. [12]. Mentioned papers describe creating of cyclograms in the cyclically organized production. The latest papers that introduce DHSP also vary in details. Some limit the problem to a single hoist Hindi and Fleszar [2] or do not consider workstation groups Jégou and Kim [3]. However, all papers propose some non-exact methods—heuristics. This is understandable, because Levner and [11] proved that HSP is NP-hard type of decisive problems and calculation effort is too big for the real-time calculations in DHSP.

The Fig. 46.1 presents a (rather small) typical production line with two hoists, eight workstations, a few items during processing, one picked up by the hoist, and a few waiting in the processing queue.

In the cyclic production, an item must be available in a loading station every given interval of time. You cannot immediately change the item type. Due to the cyclic production, a line gradually is loaded with items and after a few cycles the line is full and produces one item each cycle. If we want to change the produced item type, the line must be unloaded—no item is introduced at the loading station and after the same number of cycles the line is empty, and we can start producing new item type. If we are interested in producing many item types together, we have to accept such an expensive line loading unloading phase. We may also choose order driven production.

We are going to present a solution for the dynamic problem, supporting multiple hoists, many types of items, workstation groups, multifunctional baths, and arbitrary processing sequences.

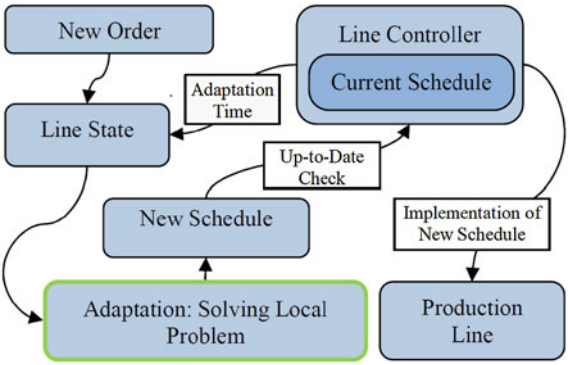


Fig. 46.2 The scheduling system

46.2.1 Scheduling System

The DHSP occurs when the production is performed as described above, but new to-be-produced items availability is not known in advance. We need to maintain a schedule, which includes currently processed items and newly ordered items. Creation of such schedules is a real-time problem because we cannot halt the production at the time when a new order is specified. If we stopped, the items that are currently processed would break the quality constraints and be destroyed.

Therefore, we propose the scheduling system (Fig. 46.2), which bases on the old schedule until it finds a feasible update.

When a new order is stated at moment τ, the line controller decides about adaptation moment t—time in the future, from which the schedule changes will be carried on $t = \tau + \Delta$, $\Delta > 0$. Line controller provides also information about the line state at the adaptation time according to current schedule. Line state is a parameter to so-called *local problem*. The scheduling algorithm generates a feasible schedule. Solving the *local problem* lasts σ amount of computation time and produces an updated schedule. Besides other constraints, the updated schedule must be the same as the current schedule up to time t. At this point, the new schedule is checked against its applicability ($\sigma < \Delta$). In case the updated schedule is infeasible, the scheduling system figures out a new adaptation time and a new local problem is solved. When the adaptation is successful, the line controller implements the new schedule to the production line. The line controller converts the schedule to some automaton language e.g. step 7.

46.3 Notion

The problem of finding the schedule for specified line, queue of items, and line state can be described as optimization problem. The problem parameters are: properties of the line (workstations locations, sizes, workstation counts), hoists

count, hoists speed. Other parameters concern produced product types: sequence of workstation for each product type, minimum and maximum times of processing in visited workstations. Additionally, parameters describing the line state in adaptation time: hoist positions, processing stages of products processed, a queue of to be processed items are also considered. A result schedule can be described by decisive variables of this problem: routes of hoists, assignment of pickup operations to hoists, assignment of put down operations to workstations in groups, times of pickup and put down operations. The optimization criterion is minimizing the time of a last put down operation. Such criterion maximizes the performance of the production line for a given product order. Similarly, to other scheduling problems, DHSP has many constraints. It makes finding any feasible solution very hard. Schedule must fulfill process requirements—processing sequence must be correct, processing times must not be shorter than specified minimum times and longer than specified maximum times. A hoist can carry only one item at the time, a workstation can process only one item at the time. Hoists cannot move outside their physical capabilities, and they cannot collide. We provide only relevant symbols, i.e. workstation groups are omitted. Extended notion is available in Kujawski [6].

Parameters are N—product types count, $n \in \{1, \ldots, N\}$. $Z = \{z_1, \ldots, z_K\} = \{o_1, \ldots, o_{KC}\} \cup \{\delta_1, \ldots, \delta_{KQ}\}$, the queue of items to be produced. $z_k \in N$—the type of item in queue's k-th place. $o_k \in N$—products, which are already processed during adaptation. $\delta_k \in N$—products from new order, $K = KC + KQ$. H—the number of hoists present on line, $h \in \{1, \ldots, H\}$. $O_n \in \{w_1, \ldots, w_{i(n)}, \ldots, w_{I(n)}\}$—the sequence of workstation group types necessary to manufacture the product of type n. $w_{i(n)}$—the processing workstation group type of an i-th stage of a product type n. $i(n) \in \{1, \ldots I(n)\}$ where $I(n)$ is the number of steps in of product type n processing sequence. Decisive variables are defined as follows: $U(h, \lambda)$, $\lambda \in \{1, \ldots, \Upsilon\}$—routes of hoists, represented as a position of the hoist h in production time λ. $\overline{t}_{k,i(z(k))}$, $\underline{t}_{k,i(z(k))}$—moments of a pick and put of i-th stage of a k-th product from a queue Z. $\overline{h}_{k,i(z(k))}$, $\underline{h}_{k,i(z(k))}$—numbers of hoists, which perform an i-th pick and put of a k-th product from the queue Z. The optimization criterion is specified as: $Q = \max\left(\underline{t}_{k,I(z(k))}\right)$—the time of last put down of a last produced item. It is the moment when the line can stop, as none of its resources is used. Optimal solution is found when the criterion reaches minimum.

Let S be the schedule for some queue Z. The schedule S is defined by having all decisive variables values set, so S sets certain values to all decisive variables of the problem—is the solution. Let us mark $\Psi(S, \tau)$ as the shifting operation. It shifts the schedule S by τ seconds. The operation changes the schedule variables as defined: $\overline{t}_{k,i(z(k))} = \overline{t}_{k,i(z(k))} + \tau$, $\underline{t}_{k,i(z(k))} = \underline{t}_{k,i(z(k))} + \tau$,

$$U(h,t) = \begin{cases} U(h, t-\tau), & t - \tau > 0 \\ U(h, 0), & t - \tau \leq 0 \end{cases}$$

Segmenting operation $\Xi(Z)$, divides products from orders to segments of items of the same type, such that: $\{\delta_1, \ldots, \delta_{KQ}\} = \{s_{1,1}, \ldots, s_{1,S_1}\} \cup \cdots \cup \{s_{\mathrm{SEG},1}, \ldots, s_{\mathrm{SEG},S_{\mathrm{SEG}}}\}$ where $\sum_{i=1,\ldots,\mathrm{SEG}} S_i = KQ$ and $s_{i,j} = s_{i,m} \forall j, m = 1, \ldots, S_i, i = 1, \ldots,$ SEG. Segmenting divides the new order to sub-sequences of items of the same time. For example for queue $Z = \{1, 1, 3, 1, 2, 2, 3, 3\}$ operation $\Xi(Z)$ means that following segments are created: {1, 1}, {3}, {1}, {2, 2}, {3, 3}.

It is important to introduce the idea of cyclogram, which is a cornerstone of cyclic production. Cyclogram is a specific type of schedule used in electroplating lines. In Lamothe et al. [8] authors claim that the cyclic production causes schedule to be periodic. Cyclogram is a schedule that represents one period, a cycle of periodic schedule. Cyclogram contains a constant number of hoist operations. Repeated over and over again, it allows production of any number of items. Cyclograms are built this way that one item is introduced and completed during one cycle. Main features of cyclogram are a capacity and a cycle time. The cycle time is a length of cyclogram in a time domain. The capacity is a number of cycles needed to be performed in order to load fully a production line and produce a first item. Afterwards, a line produces one new item each cycle. There can be many feasible cyclograms for one item type.

Let us assume that there is a number cyclograms for each item type $n \in \{1, \ldots, N\}$ referred as $V_1, \ldots, V_N$, $v_n \in \{1, \ldots, V_n\}$. $T(v_n)$—cycle-time of cyclogram v_n, $G(v_n)$—capacity of cyclogram v_n. Cyclogram can be defined similarly as schedules in DHSP—routes, times of operations and assignment of operations to hoists. $U(v_n, h, \Lambda)$, $\Lambda \in \{0, \ldots, T(v_n)\}$—routes of hoists, represented as position of the hoist h in production time Λ for cyclogram v_n. $\overline{t}(v_n, i(n))$, $\underline{t}(v_n, i(n))$—moments of pick and put of i-th stage of product n-th in cyclogram v_n, $\overline{h}(v_n, i(n))$, $\underline{h}(v_n, i(n))$—indexes of hoists which, perform a i-th pick and put of product n-th in cyclogram v_n.

Let us mark $C(x, v_n)$ as the cyclogram unfold operation, for x items of type n using cyclogram v_n. As a result of $C(x, v_n)$ we get a schedule based on cyclogram v_n, which produces x items. Cyclogram operations are repeated a number of times until x items leave the line. A schedule length $|C(x, v_n)| = (\mathrm{G}(v_n) + x) \cdot T(v_n) - \mathrm{T}_{\mathrm{load}}[v_n] + \mathrm{T}_{\mathrm{unload}}[v_n]$. $C(x, v_n)$ creates a solution for $Z = \{z_1, \ldots, z_x\}$, where $z_k = n$ by assignment:

$$
\begin{aligned}
&U(h,t) = U(v_n, h, t \bmod T(v_n)), t \in \{0, \ldots, |C(x, v_n)|\},\\
&\overline{t}_{k,i(n)} = \overline{t}(v_n, i(n)) + (k + \theta(i(n))) \cdot T(v_n), \underline{t}_{k,i(n)} = \underline{t}(v_n, i(n)) + (k + \theta(i(n))) \cdot T(v_n),\\
&\overline{h}_{k,i(z(k))} = \overline{h}(v_n, i(n)), \overline{h}_{k,i(z(k))} = \overline{h}(v_n, i(n))\\
&\theta(i(n)) = \sum_{a=1}^{a=i(n)} \begin{cases} 1 - \overline{t}(v_n, a) < \underline{t}(v_n, a) \\ 0 - \overline{t}(v_n, a) > \underline{t}(v_n, a) \end{cases}
\end{aligned}
$$

where θ is the counter of passing product between two cycles.

46.4 Scenario Selection

The most difficult part in the scheduling system is solving the local problem. This is the true stage where scheduling takes place. We use the *scenario selection* method from Kujawski [7] as a basic scheduling algorithm. The algorithm pseudo-code:

```
Scenario Selection
 1) Calculate Ξ for queue Z
 2) Result = CURRENT_SCHEDULE
 3) for x = 1 to SEG in {s_{1,1},...,s_{1,S_1}} ∪ ... ∪ {s_{SEG,1},...,s_{SEG,S_SEG}}
 3a)   Select v_{s_{x,1}} using ℑ_{s_{x,1}} and Result.
       If classification is inconclusive select
       arbitrary v_{s_{x,1}} from {1,...,V_{s_{x,1}}}
 3b)   Calculate S(x) = C(S_x, v_{s_{x,1}}).
 3c)   Find smallest τ that there exists
       collision-less routing for Result ∪ Ψ(S(x), τ).
 3d)   Result=Result ∪ Ψ(S(x), τ)
```

The scheduling system provides necessary parameters: queue Z, adaptation time t current schedule (point 2), and all the other line specific properties. We prepare cyclogram sets for each type of item before the production starts. Cyclograms can be generated by any cyclic scheduling method available as well as human created schedules. We use cyclograms generated by the method described in Kujawski [5], because besides their high performance, they support multiple-hoists, workstation groups, multifunctional tanks, etc. Point 3b) incorporates the routing algorithm. It is the same routing algorithm as described in Kujawski [5]. In the first iteration, the outermost hoist is forced to drive to the outermost workstation. Next hoists are tracing the previous hoist, staying as close to it as it is possible, while avoiding collisions. In second iteration, we begin from the last hoist and reduce all futile movements. Algorithm is minimizing the hoist route length by using the "central" workstation of the hoist zone to wait when not colliding. Then the post-processing stage eliminates any sequences of hoist movements, which can be replaced by a simple "ride and wait" or "wait and ride" scheme. We use both a constant speed and an acceleration based hoist movement models. An acceleration based model, where a hoist position is calculated with usage of an acceleration ratio, braking ratio and maximum speed is required in real-life applications. In research, like in other papers, we use the constant speed model, where the hoist moves always in maximum speed and can immediately gain such speed and break to stop. Movement models are important in a creation of valid routes for hoists.

The *scenario selection* scheduling method assumes that someone experienced in a production process is able to spot regularity in a production state. For example, "it is the best to use particular production method in case we have

produced many brake pads, and now we start to produce ball bearings". We are trying to gain from domain knowledge in order to increase throughput of the production line. This has its reason. A production line is usually designed once to produce certain types of products. Its design does not change frequently to produce something else. Production orders also may tend to be repeated, as the same customers are ordering.

In DHSP, there are many areas where such assumption would apply. We are interested in point 3a). A cyclogram is a scenario that we use to unfold a schedule. We can pick one of many cyclograms present. It is possible to choose the best scenario basing on a situation. This is a typical classification problem. We compare few popular classification algorithms: the feed-forward, back-propagation artificial neural network (ANN), naive Bayes classifier (NB), support vector machine classifier (SVM) in this paper research and calculation of presented results.

For each product type $n \in \{1, \ldots, N\}$, we construct a classifier $\Im_n$. The classifier selects a scenario used for unfolding and shifting operations during scheduling. A number of problem features are calculated to reflect the state of a line and the characteristic of the product order. Features include utilization of workstations, utilization of hoists, a number of products in segment, a length of unfolded segment, a resource collision count. These features are normalized and provided as inputs to the classifier. Classifiers are taught using randomly generated patterns.

46.5 Electroplating Line Case Analysis

The production line in Wrocław, Poland is a typical small electroplating line. It is used in production of metal furniture elements. The line consists of 16 workstation groups composed into one column. One group has three workstations, and the other groups are in fact, a singular workstation. Two hoists are available. Hoists maximum speed is 0.7 m/s. Hoist collision zone is 2 m—two hoists centers cannot be closer than 2 m or the collision occurs.

We will consider two item types that are frequently ordered together. Let us mark the technology process of technical chroming as a "type A" and process of nickeling as a "type B". Tables 46.1 and 46.2 define the processes sequence and quality constraints for process, A and B respectfully. All pick up and put down times are 8 s.

The step number 10, "Chroming" lasts extensively longer than other steps. For that reason, the chroming bath is multiplied to three workstations. All three workstations can be used in the processing.

Since we have defined the technological processes, it is now useful to introduce utilization ratio. Let us mark base solution length for item type n:

Table 46.1 Chroming process definition

No. of Steps	Step name	Min. time	Max. time	Group no.
1	Loading station	–	–	1
2	Chem. degreasing	300	420	10
3	Dripping	30	90	11
4	Electrochemical degreasing	60	180	12
5	Warm rinse	30	90	13
6	Rinse I	20	80	14
7	Cascade rinse I	20	80	15
8	Cascade rinse II	20	80	16
9	Anode etching	150	210	9
10	Chroming	1,200	1,300	8(3)
11	Salvaging rinse	20	80	7
12	Rinse II	20	80	6
13	Rinse with chrome reduction	20	80	5
14	Rinse III	20	80	4
15	Timed rinse	20	80	3
16	Blow in bath	30	90	2
17	Unloading station	–	–	1

Table 46.2 Nickeling process definition

No. of Steps	Step name	Min. time	Max. time	Group no.
1	Loading station	–	–	1
2	Timed rinse	300	400	3
3	Salvaging rinse I	30	90	5
4	Salvaging rinse II	30	90	7
5	Degreasing I	60	180	10
6	Electrochemical degreasing	20	80	12
7	Rinse I	20	80	14
8	Cascade rinse I	20	80	16
9	Cascade rinse II	20	80	15
10	Warm rinse	150	210	13
11	Nickeling	200	400	11
12	Degreasing II	20	80	10
13	Rinse II	20	80	6
14	Rinse III	20	80	4
15	Blow in bath	30	90	2
16	Unloading station	–	–	1

$$TB_n = \bigcup_{i \in I^n} (m_{i,n}) + \bigcup_{0 \le i \le I(n)-1} \left(\left\lceil \frac{|P_i - P_{i+1}|}{v} \right\rceil + t_{i,n}^{up} \right) + \bigcup_{1 \le i \le I(n)} \left(t_{i,n}^{down} \right)$$

where $m_{i,n}$ is minimum processing time of product n in stage i, P_i is a position of workstation performing i-th stage, $t_{i,n}^{up}$ and $t_{i,n}^{down}$ are times of picking and putting

Table 46.3 The workstation centers positions

Group no.	1	2	3	4	5	6	7	8(1)	8(2)
Position (m)	0.3	0.94	1.56	2.18	2.8	3.43	4.08	4.85	5.75
Group no.	8(3)	9	10	11	12	13	14	15	16
Position (m)	6.65	7.46	8.14	8.76	9.42	10.08	10.7	11.27	11.79

item to workstation respectively. Base solution length is ratio at which items are produced in naive schedule, where each stage of item production takes the minimum time and is immediately transferred to workstation of next stage and next item is introduced after previous finishes. Utilization ratio of schedule S for some queue Z is defined as follows:

$$U^{ratio}(S,Z) = \frac{\bigcup_{0 \le k \le K} TB_{z_k}}{|S|},$$

where $|S|$ is length of schedule S.

Utilization ratio shows how successful is scheduling procedure. If $U^{ratio}(S,Z) < 1$ the procedure is worse than using naive schedule. If $U^{ratio}(S,Z) \ge 2$ the created schedule is able to produce items at least twice as fast as naive schedule. It allows comparing two scheduling methods more reliably because it is only slightly dependent to queue size.

To calculate the time required to move a hoist from the workstation, i to j the distance between workstation centers is divided by the maximum hoist speed and rounded up to whole seconds. In order to transport an item from a group hoist need to go to the group center, pick the item up (8 s), move to the destination bath and put the item down (8 s) (Table 46.3).

46.5.1 Production

The organization on the Wrocław line is currently cyclical. The owners want to switch to the dynamic production, because they need to produce many item types together. As a data to research, we will use some historical runs and typical order ratios. For historical runs the schedules are available but without the real-time aspect. We can schedule offline as we know the whole order before production is started. Such test is also valuable because it allows benchmarking the performance of created schedules (Table 46.4).

More general production requirements, which we have on the Wrocław line, is that statistically orders are composed like 1:2, so for each item of type Two items of the type B are ordered. We are going to analyze the ratio 2:3 and 1:1.87. For each ratio, the orders are generated. We are going to simulate the real-time production. We generate the queue of 40 items for each proportion (16 × A:24 × B and 18 × A:22 × B) with random sequence. We calculate the schedule for three cases: whole order is known at the beginning, the next segment is known in 100 s before the shortest available

Table 46.4 Queues to analyze

Queue	Items sequence
Queue 1	AABBAABB
Queue 2	7 × A; 5 × B; 5 × A
Queue 3	20 × A;15 × B
Queue 4	5 × A; 5 × B; 5 × A; 5 × B; 5 × A
Queue 5	12 × B; 4 × A; 6 × B; 2 × A
Queue 6	8 × B; 8 × A

adaptation time, and the next segment is known in the time of the shortest adaptation time. The shortest adaptation time is the time where the last item of current schedule is loaded to line, adaptation time must be greater to keep the correct sequence of item introduction. The result is averaged from 100 generated instances.

46.5.2 Algorithm Work Example

We have prepared a V_1, V_2 cyclograms using Kujawski [5]. Initially there were 56 and 30 different cyclograms. During preparation of patterns for classifier learning, we picked the most significant 3 cyclograms of each type. We have continued with cyclograms for item type A: A1—length 415 s., capacity 7; A2—length 424 s., capacity 6; A3—length 432 s., capacity 6; and for type B: B1—length 240 s., capacity 6; B2—length 257 s., capacity 5; B3—length 532 s., capacity 3. We have trained two neural networks, each for the different type of item.

Let us analyze the *scenario selection* method for Queue 1. The classifiers $\Im_n$ are neural networks in this case. Let us assume that we are starting the production, so the line and schedule are empty. Ξ operation gives us four segments {AA}, {BB}, {AA}, {BB}. In the first step $\Im_1$ returns A3, $S(x)$ is a production schedule of two items of type A, its length is 2,887 s. $\tau = 0$ as the current schedule is empty, so there cannot be any collisions. We store the result and proceed to a second iteration. $\Im_2$ returns B3, We calculate S(x), its length is 1,720 s. We analyze $\tau > 440$ s and find $\tau = 893$ constructs a feasible schedule for AABB with a length of 2,887 s. In a third iteration $\Im_1$ returns A2, We calculate $S(x)$ by cyclogram unfolding, its length is 2,967 s. We analyze $\tau > 1{,}433$ s and find $\tau = 2{,}464$. We store the result (5,431 s.) and proceed to the last iteration. $\Im_2$ returns again B3, We calculate $S(x)$, its length is 1,720 s. We analyze $\tau > 2{,}896$ s and find $\tau = 3{,}424$. The schedule for Queue 1 is created and its length is 1:30:31 (5,431 s.). Scheduling takes around 3 s. All calculations are performed on Intel Q9300 2.5 GHz processor.

46.5.3 Results

The tables summarize the results of the *scenario selection* method (Tables 46.5, 46.6, 46.7 and 46.8).

Table 46.5 Historical queues results

	Queue 1	Queue 2	Queue 3	Queue 4	Queue 5	Queue 6
Historical length	02:41:36	03:28:48	04:34:51	05:11:12	04:02:48	02:33:24
His. utilization	1.37	1.92	3.34	2.06	2.97	2.88
ANN length	01:30:31	02:34:18	03:59:49	04:13:57	03:41:36	02:20:56
ANN utilization	2.44	2.60	3.83	2.53	3.52	3.14
ANN improvement (%)	43.89	26.10	12.75	18.40	15.68	8.13
NB length	02:32:26	02:35:18	04:00:20	04:16:24	03:46:41	02:20:56
NB utilization	1.45	2.59	3.82	2.50	3.44	3.14
NB improvement (%)	5.67	25.62	12.56	17.61	13.74	8.13
SVM length	01:53:43	02:50:19	04:00:13	04:43:07	03:42:43	02:52:25
SVM utilization	1.94	2.36	3.82	2.27	3.50	2.56
SVM improvement (%)	29.63	18.43	12.60	9.02	15.25	−12.40

Table 46.6 Ratio 16:24; The scenario selection

Test type	Avg. schedule length	Avg. utilization ratio	Hoists utilization (%)	Baths utilization (%)
ANN entire order known	09:24:32 ± 00:28:46	2.09	21.98	8.47
NB entire order known	10:31:00 ± 00:38:15	1.87	20.59	8.24
SVM entire order known	09:34:42 ± 00:30:01	2.06	19.08	8.18
ANN 100 s before	09:23:07 ± 00:29:05	2.10	22.04	8.50
NB 100 s before	10:32:54 ± 00:38:13	1.87	20.53	8.21
SVM 100 s before	09:34:08 ± 00:30:13	2.06	19.10	8.19
ANN adaptation time	09:40:37 ± 01:09:02	2.04	21.34	8.26
NB adaptation time	10:27:11 ± 00:37:34	1.89	20.61	8.27
SVM adaptation time	09:27:08 ± 00:28:33	2.08	19.24	8.26

Table 46.7 VII ratio 18:22; The scenario selection

Test type	Avg. schedule length	Utilization ratio	Hoists utilization (%)	Baths utilization (%)
ANN entire order known	9:33:57 ± 1:00:36	2.00 ± 0.18	21.06	8.52
ANN 100 s before	10:44:49 ± 0:12:30	1.77 ± 0.03	20.14	8.22
ANN adaptation time	10:44:00 ± 0:12:23	1.77 ± 0.03	18.88	8.21

Table 46.8 Ratio 16:24; Different queue sizes

	Avg. schedule length	Utilization ratio	Hoists utilization (%)	Baths utilization (%)
20 items	05:05:03 ± 00:35:03	1.96 ± 0.03	20.14	8.52
30 items	07:35:11 ± 00:54:07	1.97 ± 0.09	20.43	8.44
40 items	09:34:02 ± 00:53:33	2.07 ± 0.08	21.51	8.21
50 items	13:26:27 ± 00:44:31	1.87 ± 0.05	20.32	9.08

Comparison between human created historic queues and our results shows 8 up to 29% improvement in all cases.

For randomly created queues, where scheduling is required to be maximally flexible, the utilization ratio is still higher than 2.0, and shows, that statistically the line allows production of at least two products at time. We compared the different classifiers performance only on one production ratio, but it seems that neural network and support vector machines classifiers are more effective than naive Bayes classifier. We also can see that both hoist and bath utilization (ratio of time busy to total length of schedule) does not seem to have any influence on throughput criterion.

Four performed tests show that utilization ratio allows comparing results of the schedules for queues of the different length. While length of the schedule is increasing with increasing number of processed products, the utilization ratio stays at more less the same level.

The scheduling algorithm is efficient enough for the Wrocław line and possibly for similar sized productions. During the tests, there were no cases, where the new schedule was out of date and required new adaptation time.

References

1. Frohlich R, Steneberg SC (2009) Optimal cyclic multiple hoist scheduling for processes with loops and parallel resources. IEEE Int Conf Syst Man Cybern, SMC 2009
2. Hindi KS, Fleszar K (2004) A constraint propagation heuristic for the single-hoist, multiple-products scheduling problem. Comput Ind Eng 47(1):91–101
3. Jégou D, Kim DW (2006) A contract net based intelligent agent system for solving the reactive hoist scheduling problem. Expert Syst Appl 30(2):156–167
4. Jiyin L, Yun J (2002) Cyclic scheduling of a single hoist in extended electroplating lines: a comprehensive integer programming solution. IIE Trans 34(10):905
5. Kujawski K (2006) Zagadnienie optymalizacji działania wybranych linii produkcyjnych. Wydział Informatyki i Zarządzania. Wrocław, Polska, Politechnka Wrocławska. Mgr inż.: 79
6. Kujawski K, Świątek J (2009) Using cyclogram unfolding in dynamic hoist scheduling problem. In: Proceedings of the international conference on systems engineering ICSE'09
7. Kujawski K, Świątek J (2010) Intelligent scenario selection in dynamic hoist scheduling problem: the real-life electroplating production line case analysis. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July, 2010, London, UK
8. Lamothe J, Correge M, et al (1995) A dynamic heuristic for the real time hoist scheduling problem. In: Proceedings of ETFA '95, Symposium on emerging technologies and factory automation 1995, INRIA/IEEE 1995
9. Lamothe J, Thierry C et al (1996) A multihoist model for the real time hoist scheduling problem. Engineering in Systems Application Multiconference, Lille 461–466
10. Leung J, Guoqing Z (2003) Optimal cyclic scheduling for printed circuit board production lines with multiple hoists and general processing sequence. IEEE Trans Robot Autom 19(3):480–484
11. Levner E, Kats V et al (2010) Complexity of cyclic scheduling problems: a state-of-the-art survey. Comput Ind Eng **59**(2): 352–361
12. Mak RWT, Gupta SM et al (2002) Modeling of material handling hoist operations in a PCB manufacturing facility. J Electr Manuf 11(1):33

13. Phillips LW, Unger PS (1976) Mathematical programming solution of a hoist scheduling program. AIIE Trans 8(2):219–225
14. Yan PC, Chu et al (2008) An algorithm for optimal cyclic scheduling in a robotic cell with flexible processing times. IEEE Int Conf Ind Eng Eng Manag, IEEM 2008
15. Zhou Z, Liu J (2008) A heuristic algorithm for the two-hoist cyclic scheduling problem with overlapping hoist coverage ranges. IIE Trans 40(8):782–794

Chapter 47
Effect of HAART on CTL Mediated Immune Cells: An Optimal Control Theoretic Approach

Priti Kumar Roy and Amar Nath Chatterjee

Abstract Highly active antiretroviral therapy (HAART) reduces the virus load during long term drug therapy. It is observed that during drug therapy virus load sustained in the immune system also CTL is generated due to activation of immune cells and declination of memory CTL which is actually the causal effect of suppression of virus load. Here we extended our work (Roy PK, Chatterjee AN (2010) Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, pp 615–620) and formulate a set of differential equations to study the effect of HAART on immune cells to a HIV infected individuals. We also incorporate in our model of an optimal control strategy during drug treatment, which reduces the infected cell population and increases the uninfected cell population. It is to be mentioned here that the control variable is used as drug dose which introduces in the diseases transmission term. Analytical and numerical study shows that optimal control process can reduces the infected cell population. An objective function is also introduced to minimize the systemic cost of chemotherapy.

47.1 Introduction

Over the last two decades there has been an extensive effort to make in the mathematical modelling of HIV, representing the virus that causes Acquired Immune Deficiency Syndrome or AIDS. Human Immunodeficiency Virus (HIV)

P. K. Roy (✉) · A. N. Chatterjee
Centre for Mathematical Biology and Ecology, Department of Mathematics, Jadavpur University, Kolkata, 700032, India
e-mail: pritiju@gmail.com

A. N. Chatterjee
e-mail: anchaterji@gmail.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_47,
© Springer Science+Business Media B.V. 2011

targets the immune cells mainly CD4 positive T lymphocytes ($CD4^+$T cells, a type of white blood cells), which is the main component of immune system. When the number of $CD4^+$T cells is reduced below 200 mm^{-3} the HIV infected patients is treated as AIDS patients [9, 10]. In this research article our goal has been generated to formulate a mathematical model consisting a set of differential equations which describes the interaction of HIV and the immune system for the purpose of specific antiviral treatment strategies and the exploration of optimal control of this model equation.

In recent times essentially there are using two types of anti-HIV drugs reduced into two classes: reverse transcriptase inhibitors (RTI) which prevent HIV from infecting cells or put a stop to go infection of new cells, whereas protease inhibitors (PIs) prevent with the production of new infectious virons by infected cells [13, 14] causing the virus to be unable to infect helper-T cells. Moreover it is to be noted here that these two or more types of drugs can be used as HAART and hence reduce the viral load from the infected individuals, also increases $CD4^+$T cells count. It is also been observed that during long term treatment of HIV infected patient by using HAART give results 10–100 fold reduction of virus load and 25% increases in $CD4^+$T cells and CTL count. HAART containing six RT inhibitor (AZT, ddI, d4T, 3TC, and zevirapine) and three P.I inhibitor (saquinavir, indinavir, and ritonavir) which is administered for many patients [5].

HIV grows weaker in the immune system due to its infection, but the virus still works inside the immune system. Here role of the antigen-presenting cell (APC) of a HIV infected individual is thus important since it indicate precursor Cytotoxic T Lymphocytes to differentiate into killer T-cells which is known as effectors CTL [1]. On the other hand killer T-cells instigate to destroy infected $CD4^+$T cells from where new infectious virons is born. This imply that there has been several complications in the immune system for using drugs or more precisely it can be say that the infection process is quite complicated interaction process between different cells, virus and drug [12]. In this paper we used to investigate how specific antiviral treatment can affect the immune response i.e. whether this treatment can predominantly reduce the viral load and in another sense how it control the disease progression in a long-term treatment of HIV infected patients.

Mathematical models of drug treatment dynamics suggest that the CTL response could be maintained or even increased by combining of drug therapy with vaccination [13, 14]. When drug is administered in a HIV infected individual CTL is stimulated and it acts against the infected $CD4^+$T cells. It is also been observed that in a long term use of these highly cost drug therapy which give results of many complication. After taking these drugs patient may be suffered by harmful pharmaceutical side effects such as cardiovascular, lactic acidosis, etc. [7]. Now a day's many clinical laboratories keep a systematic data records of patient treatment courses with respect to effectiveness and results. But that records provide incompatible indication as to which is better: early treatment (defined as $CD4^+$T cell counts between $200-500\,mm^{-3}$ of blood) or treatment at a later stage (below 200 mm^{-3}). "Better" here is based on overall health of

patient (i.e., side effects) and a preservation or amplify in the $CD4^+T$ cells count [6]. Thus it is imperative to say that effective HAART use is oppressive for some patients and impossible to other peoples due to higher costs of drug and for complicated course of drug therapy. To avoid such complications of the results, we also performed a further analysis to investigate under that circumstances introducing a control variable in the proposed model where the control variable is actually the drug dose. This control affects the interaction between infected and uninfected $CD4^+T$ cells and the cost function is to be maximized the uninfected T cell, CTL population and to minimized the infected T cell population and accumulated side effect.

In Sect. 47.2 we present a working model which is an extended work of [11] and also based upon CTL responses due to antiviral treatment. In Sect. 47.3 we present the analytical study of the model where we find four equilibria with their existence condition. Here we find out their local stability property. In Sect. 47.4 we present the optimal control problem in which the contact process between he uninfected and infected T cell is controlled. In this section we find out the optimal control variable for which the uninfected $CD4^+T$ cell and CTL levels systematic cost of chemotherapy. The optimal control is characterized by use of Pontryagin's Minimum Principle [4]. We also find out the condition for which the system is unique. In Sect. 47.5 we solve the model numerically. In Sect. 47.6 we discuses the analytical and numerical results according to biological aspect.

47.2 Presentation of Mathematical Model

To establish a mathematical model of immune cell infection by HIV, we first consider the basic dynamics of T cell in absence of HIV. T cell is produced in the bone marrow and these immature cells migrate to thymus and they are matured to immunocompetent T cells. These T cells can be created by the proliferation of existing T cells [9, 10]. In absence of HIV the model of T cell dynamics is [9]

$$\dot{x} = \lambda + px\left(1 - \frac{x}{T_m}\right) - d_1 x \tag{47.1}$$

where x is the number of $CD4^+T$ cells as measured in blood, λ is the constant rate of supply of uninfected T cell from precursors in the thymus, p is the maximum proliferation rate and T_m is the T cell population density at which proliferation shuts off. Since T cell, like all other cell have a natural life span thus we consider d_1 which represents the natural death rate constant of uninfected $CD4^+T$ cells.

Under this circumstances there are certain parametric restrictions on the basis of realistic population dynamics specially in the cell biological system. We shall assume that $\lambda > 0$ as because thymus remains functional. The steady state

population of x always less than T_m, so that T cell population will expand when stimulated during the process of virus activity. Thus we choose $d_1 T_m > \lambda$ [9].

In the presence of HIV infection the uninfected CD4$^+$T cell becomes infected by free virus at a rate $\beta_1 xy$, where y is the number of infected CD4$^+$T cells and β_1 is the rate of infection of uninfected T cell.

Then the model for virus dynamics is

$$\begin{aligned} \dot{x} &= \lambda + px\left(1 - \frac{x}{T_m}\right) - d_1 x - \beta_1 xy \\ \dot{y} &= \beta_1 xy - d_2 y \end{aligned} \tag{47.2}$$

where d_2 is the death rate of virus producing cell i.e. infected CD4$^+$T cell. Since we are interested to see the effects of immune responses in the long term drug therapy [2], therefore we include another variable z_d (i.e. drug induced CTL) responses against virus infected cells. Thus the modified model becomes

$$\begin{aligned} \dot{x} &= \lambda + px\left(1 - \frac{x}{T_m}\right) - d_1 x - \beta_1 xy \\ \dot{y} &= \beta_1 xy - d_2 y - \beta_2 y z_d \\ \dot{z_d} &= sy - d_3 z_d \end{aligned} \tag{47.3}$$

where β_2 is the killing rate of virus producing cell by CTL, s is the rate of stimulation of drug induced CTL and d_3 is the natural death rate of drug induced CTL. It is to be noted here that, our immune system induces CTL responses through the interactions with dendritic cells and we think about that CTL responses z_e include both CTL effector and memory CTL [3]. Again, in the absence of immune impairment effect, proliferation of CTL depends upon the host population and viral replication together with the initial conditions [3, 13]. Thus in a nutshell proliferation of this CTL is described by $cxyz_e$, where c is the proliferation rate, d_4 is the rate of decay and β_3 is the killing rate of virus producing cell by z_e.

By the above mentioned assumptions, we can write down the following set of differential equations under HIV induced immunological system as

$$\begin{aligned} \dot{x} &= \lambda + px\left(1 - \frac{x}{T_m}\right) - d_1 x - \beta_1 xy \\ \dot{y} &= \beta_1 xy - d_2 y - \beta_2 y z_d - \beta_3 y z_e \\ \dot{z_d} &= sy - d_3 z_d \\ \dot{z_e} &= cxyz_e - d_4 z_e \end{aligned} \tag{47.4}$$

The above system has to be analyzed with the following initial condition: $\{x(0) > 0,\ y(0) > 0,\ z_d(0) > 0,$ and $z_e(0) > 0\}$. We observe that the right hand side of the above function is a smooth function of the variable (x, y, z_d and z_e) and the parameters with their non-negative condition. So local existence and uniqueness properties holds in R^4_+.

47.3 Equilibria

We find that the system has following equilibria $E_0(x_0, 0, 0, 0)$, $E_1(x_1, y_1, 0, 0)$, $E'(x', y', z_d', 0)$, and $E^*(x^*, y^*, z_d^*, z_e^*)$ where, $x_0 = \frac{T_m}{2p}[(p - d_1) + \sqrt{(p - d_1)^2 + \frac{4p\lambda}{T_m}}]$, $x_1 = \frac{d_2}{\beta_1}$, $y_1 = \lambda + \frac{d_2(p-d_1)}{\beta_1} - \frac{{d_1}^2 p}{{\beta_1}^2 T_m}$, $x' = \frac{\left(\frac{\beta_1 d_2 d_3}{\beta_2 s} - d_1 + p\right) + \sqrt{\left(\frac{\beta_1 d_2 d_3}{\beta_2 s} - d_1 + p\right)^2 + 4\lambda\left(\frac{p}{T_m} + \frac{d_1 {\beta_1}^2}{s\beta_2}\right)}}{2\left(\frac{p}{T_m} + \frac{d_1 {\beta_1}^2}{s\beta_2}\right)}$, $y' = \frac{d_3(\beta_1 x_3 - d_2)}{s\beta_2}$, $z_d' = \frac{\beta_1 x_3 - d_2}{\beta_2}$, and $x^* = \frac{T_m}{2pc}\left[c(p - d_1) + \sqrt{c^2(p - d_1)^2 + \frac{4pc}{T_m}(c\lambda - \beta_1 d_4)}\right]$, $y^* = \frac{d_4}{cx^*}$, $z_d{}^* = \frac{sd_4}{d_3 cx^*}$, $z_e{}^* = \frac{d_3 cx^*(\beta_1 x^* - d_2) - s\beta_2 d_4}{d_3 c\beta_3 x^*}$. The first equilibrium E_0 represents the uninfected equilibrium with minimum level of healthy CD4$^+$T cells and there exist no infected cell as well as CTL response (both drug induced and effector CTL). Note that the absence of CTL means immune response in presence of HIV. The second equilibrium E_1 represents the presence of uninfected and infected CD4$^+$T cells and there exists no immune response. This system is very harmful for HIV infected patients. The third equilibrium E' represents the presence of uninfected and infected T cells together with drug activated CTL. In this case there exists no effector CTL. The interior equilibrium E^* represents the presence of uninfected T-cell, infected T-cell, drug induced CTL and effector CTL.

Remark Existence condition for E_0: $p > d_1$, for E_1: $p > \frac{\beta_1 T_m(d_1 d_2 - \lambda\beta_1)}{d_2(\beta_1 T_m - d_2)}$, for E': $p > (d_1 - \frac{d_2 d_3 \beta_1}{s\beta_2})$ and for E^*: $p > \frac{\beta_1 d_1 T_m}{\beta_1 T_m - d_2}$, if $c > \frac{\beta_1 d_4}{\lambda}$. It is observed that E_1 arises for E_0 if $p = \frac{\beta_1 T_m(d_1 d_2 - \lambda\beta_1)}{d_2(\beta_1 T_m - d_2)}$ and persists for all $p > \frac{\beta_1 T_m(d_1 d_2 - \lambda\beta_1)}{d_2(\beta_1 T_m - d_2)}$. We also observe that the equilibrium E' arises for E_1 if $p = (d_1 - \frac{d_2 d_3 \beta_1}{s\beta_2})$ and persists for $p > (d_1 - \frac{d_2 d_3 \beta_1}{s\beta_2})$. For $p > \frac{\beta_1 d_1 T_m}{\beta_1 T_m - d_2}$, and $c > \frac{\beta_1 d_4}{\lambda}$, E^* become stable and E' does not exist.

47.4 Stability Analysis

In this section we establish stability results for steady state E_0, E_1, E', E^*.

Property 1 *The uninfected equilibrium E_0 of the system (47.4) is stable for $x_0 < \frac{d_1}{\beta_1}$, $p > d_1$ and for $|p| < \frac{\alpha_2' + \sqrt{\alpha_2'^2 - 4\alpha_1'\alpha_3'}}{2\alpha_1'}$, where, $\alpha_1' = 1 - (T_m - \frac{2d_2}{\beta_1})^2$, $\alpha_2' = 2d_1 - \frac{4\lambda}{T_m} + \frac{4d_1 d_2 T_m}{\beta_1} - 2{T_m}^2 d_1$, and $\alpha_3' = {d_1}^2(1 - {T_m}^2)$.*

Proof The characteristic equation of the linearized system of (47.4), corresponding to E_0, is given by $\{\rho - p(1 - \frac{2x_0}{T_m}) - d_1\}(\rho - \beta_1 x_0 + d_2)(\rho + d_3)(\rho + d_4) = 0$. The eigen values are $p(1 - \frac{2x_0}{T_m}) - d_1, \beta_1 x_0 - d_2, -d_3$, and $-d_4$. Since one eigen value is negative (i.e. $\rho = -d_4 < 0$), the local stability of E_0 demands the negative real parts of all roots of the equation. Which gives

$x_0 < \frac{d_2}{\beta_1}$, and $|p| < \frac{\alpha_2' + \sqrt{\alpha'^2 - 4\alpha_1\alpha_3}}{2\alpha_1}$, where $\alpha_1' = 1 - (T_m - \frac{2d_2}{\beta_1})^2$, $\alpha_2' = 2d_1 - \frac{4\lambda}{T_m} + \frac{4d_1d_2T_m}{\beta_1} - 2{T_m}^2 d_1$, and $\alpha_3' = {d_1}^2(1 - {T_m}^2)$. For the above condition E_0 is stable and when this inequality is reversed, E_0 loses stability and the infected steady state becomes a locally asymptotically stable spiral point. Thus we can say that if the proliferation rate is restricted and the uninfected T cell population is less than $\frac{d_2}{\beta_1}$ then the system is locally stable around E_0. □

Property 2 *System* (47.4) *is locally asymptotically stable around* E_1 *if* (i) $p < d_1$ (ii) $p > \max\{(d_1 + d_2 + d_3) - \frac{s\beta_2}{2\beta_1}, -d_2 + \sqrt{{d_2}^2 + (2d_1 + 3d_2 + 2d_3)\beta_1}\}$.

Proof The characteristic equation of the linearized system of (47.4), corresponding to E_1, is given by $(\rho + d_4 - cx_2y_2)(\rho^3 + A_1\rho^2 + A_2\rho + A_3) = 0$, where $A_1 = d_1 + d_2 + d_3 + \beta_1y_1 - p + x_1(\frac{2p}{T_m} - \beta_1), A_2 = d_2d_3 + s\beta_2y_2 - d_3\beta_1x_2 + (d_2 + d_3 - \beta_1x_2)\{d_1 + \beta_1y_2 - p(1 - \frac{2x_2}{T_m})\}, A_3 = d_3{\beta_1}^2x_2y_2 + (d_2d_3 + s\beta_2y_2 - d_3\beta_1x_2)\{d_1 + \beta_1y_2 - p(1 - \frac{2x_2}{T_m})\}$. Since one eigenvalue is negative, $\rho = -(d_4 - cx_2y_2)$, thus the system is locally stable around E_1 if $\rho^3 + A_1\rho^2 + A_2\rho + A_3 = 0$ has a negative real part. Now it is obvious that according to Routh–Hurwith criteria E_1 is asymptotically stable if $p < d_1$, $p > \max\{(d_1 + d_2 + d_3) - \frac{s\beta_2}{2\beta_1}, -d_2 + \sqrt{{d_2}^2 + (2d_1 + 3d_2 + 2d_3)\beta_1}\}$. Thus we can say that if the proliferation rate is below the death rate of uninfected T cell then in the presence of infection the system is stable around E_1. □

Property 3 *System* (47.4) *is locally asymptotically stable around* E' *if* (i) $\frac{d_2{\beta_1}^2 + \sqrt{{d_2}^2{\beta_1}^4 + 4sd_2{\beta_1}^3\beta_2}}{2\beta^3} < x_3 < \min\{\frac{d_3}{\beta_1}, \frac{s\beta_2 + d_2\beta_1}{{\beta_1}^2}\}$ (ii) $s\beta_2 < \frac{{d_3}^2 - d_2d_3\beta_1}{d_2\beta_1}$.

Proof Proof is obvious.

Property 4 *The interior equilibrium* E^* *is locally asymptotically stable if* $p > \max\{\frac{\beta_1T_m}{2}, \frac{\beta_1(d_1+d_3)}{\beta_1 - \frac{4pd_2}{T_m}}, \frac{d_1\beta_1}{d_3\beta_1 + \frac{2d_2d_3}{T_m}}, \frac{cd_3{\beta_1}^2}{2sd_4\beta_2(1+\frac{1}{T_m})}\}, p < \frac{{sd_4}^2\beta_2T_m}{2pd_3c^2{x^*}^2}, c > \frac{\beta_1(\beta_1d_4 + d_3 - d_3\beta_1)}{d_3(d_2+\beta_2)}$, $\beta_3d_4 < 1$, *and* $\frac{d_4(d_2+\beta_2)}{s\beta_2+\beta_1} < x^* < \frac{d_3}{d_3+\beta_1}$.

Proof The characteristic equation of the linearized system of (47.4), corresponding to E^*, is given by $\rho^4 + C_1\rho^3 + C_2\rho^2 + C_3\rho + C_4 = 0$ where, $C_1 = \frac{\lambda}{x^*} + \frac{px^*}{T_m} + d_3, C_2 = d_3(\frac{\lambda}{x^*} + \frac{px^*}{T_m}) + {\beta_1}^2x^*y^* + s\beta_2y^* + c\beta_3y^{*2}{z_e}^*, C_3 = (\frac{\lambda}{x^*} + \frac{px^*}{T_m})(s\beta_2y^* + c\beta_3y^{*2}{z_e}^*) + (d_3 + \beta_1)c\beta_3x^*{z_e}^* + d_3\beta_1x^*y^*, C_4 = cd_3\beta_3y^{*2}{z_e}^*(\frac{\lambda}{x^*} + \frac{px^*}{T_m} + \beta_1)$, and $\Delta = C_1C_2 - C_3$. From the above characteristic equation it is obvious that $C_1 > 0, C_4 > 0$, and $\Delta > 0$. Therefore using Routh–Hurwitz criterion, we find that all roots of the characteristic equation for E^* will have negative real parts if they satisfy the condition given below $p > \max\left\{\frac{\beta_1T_m}{2}, \frac{\beta_1(d_1+d_3)}{\beta_1 - \frac{4pd_2}{T_m}}, \frac{d_1\beta_1}{d_3\beta_1 + \frac{2d_2d_3}{T_m}},\right.$

$\frac{cd_3{\beta_1}^2}{2sd_4\beta_2(1+\frac{1}{T_m})}$ $\Bigg\}$, $p < \frac{sd_4{}^2\beta_2 T_m}{2pd_3c^2x^{*2}}$, $c > \frac{\beta_1(\beta_1 d_4+d_3-d_3\beta_1)}{d_3(d_2+\beta_2)}$, $\beta_3 d_4 < 1$, and $\frac{d_4(d_2+\beta_2)}{s\beta_2+\beta_1} < x^* < \frac{d_3}{d_3+\beta_1}$.

Thus the system is asymptotically stable around the interior equilibrium E^*, if it satisfy the above condition. □

47.5 The Optimal Control Problem on Chemotherapy

In this section our main object is to minimize the infected CD4$^+$T cells population as well as minimize the systemic cost of drug treatment. In order to that we formulate an optimal control problem. We also want to maximize the level of healthy CD4$^+$T cells. So in the model (47.4) we use a control variable $u(t)$, represents the drug dose satisfying $0 \le u(t) \le 1$. Here $u(t) = 1$ represents the maximal use of chemotherapy and $u(t) = 0$ represents no treatment. We choose our control class measurable function defined on $[t_0, t_f]$ with the condition $0 \le u(t) \le 1$, i.e. $U := \{u(t) | u(t)$ is measurable, $0 \le u(t) \le 1,\ t \in [t_0, t_f]\}$. Based on the above assumption, the optimal control problem is formulated as:

$$J[u] = \int_{t_0}^{t_f} [Py(t) + Ru^2(t) - Qx(t)]\, dt \tag{47.5}$$

subject to the state system

$$\begin{aligned} \dot{x} &= \lambda + px\left(1 - \frac{x}{T_m}\right) - d_1 x - (1 - u(t))\beta_1 xy \\ \dot{y} &= (1 - u(t))\beta_1 xy - d_2 y - \beta_2 y z_d - \beta_3 y z_e \\ \dot{z_d} &= sy - d_3 z_d \\ \dot{z_e} &= cxyz_e - d_4 z_e. \end{aligned} \tag{47.6}$$

The parameter $P \ge 0$, $Q \ge 0$, and $R \ge 0$ represents the weight constant on the benefit of cost. Our main aim is to find out the optimal control variable u^* satisfying $J(u^*) = \min_{0 \le u(t) \le 1} J(u)$.

47.5.1 Existence Condition of an Optimal Control

The boundedness of the system (47.6) for the finite time interval $[t_0, t_f]$ prove the existence of an optimal control. Since the state system (47.6) has bounded coefficient, hence we can say that the set of controls and corresponding state variables is nonempty. The control set U is convex. The right hand side [8] of the state system (47.6) is bounded by a linear function. The integral of the cost functional

$Py(t) + Ru^2(t) - Qx(t)$ is clearly convex on U. Let there exist $c_1 > 0$ and $c_2 > 0$ and $\eta > 1$ satisfying $Py(t) + Ru^2(t) - Qx(t) \geq c_1|\epsilon|^\eta - c_2$. Because the state system is bounded, the existence of optimal control problem is established.

47.5.2 Characterization of an Optimal Control

Since the existence of the optimal control problem (47.5) and (47.6) established, then Pontrygin's Minimum Principle is used to derive necessary condition on the optimal control [4]. We derive the Lagrangian problem in the following form as follows

$$\begin{aligned} L(x, y, z_d, z_e, u, \xi_1, \xi_2, \xi_3, \xi_4) = {} & Py(t) + Ru^2(t) - Qx(t) \\ & + \xi_1[\lambda + px(1 - \frac{x}{T_m}) - d_1 x - (1 - u(t))\beta_1 xy] \\ & + \xi_2[(1 - u(t))\beta_1 xy - d_2 y - \beta_2 y z_d - \beta_3 y z_e] \\ & + \xi_3(sy - d_3 z_d) + \xi_4(cxyz_e - d_4 z_e) \\ & - w_1(t)u(t) - w_2(t)(1 - u(t)) \end{aligned} \tag{47.7}$$

where $w_1(t) \geq 0$ and $w_2(t) \geq 0$ are the penalty multipliers satisfying that at an optimal control $u^*(t)$, $w_1(t)u(t) = 0$, and $w_2(t)(1 - u(t)) = 0$.

The existence condition for the adjoint variable is given by

$$\begin{aligned} \dot{\xi}_1 &= -\frac{\partial L}{\partial x} = -\left[-Q + \xi_1 \left\{ p\left(1 - \frac{2x}{T_m}\right) - d_1 - (1 - u(t))\beta_1 y \right\} \right. \\ &\quad \left. + \xi_2(1 - u(t))\beta_1 y + \xi_4 c y z_e \right] \\ \dot{\xi}_2 &= -\frac{\partial L}{\partial y} = -[P - \xi_1(1 - u(t))\beta_1 x + \xi_2\{(1 - u(t))\beta_1 x - d_2 - \beta_2 z_d - \beta_3 z_e\} \\ &\quad + \xi_3 s + \xi_4 c x z_e] \\ \dot{\xi}_3 &= -\frac{\partial L}{\partial z_d} = -[-\xi_2 \beta_2 y - \xi_3 d_3] \\ \dot{\xi}_4 &= -\frac{\partial L}{\partial z_e} = -[-\xi_2 \beta_3 y + \xi_4(cxy - d_4)] \end{aligned} \tag{47.8}$$

where $\xi(t_f) = 0$ for $i = 1, 2, 3, 4$ are the transversality condition. The lagrangian is minimized at the optimal u^*, so the lagrangian with respect to u at u^* is zero. Now,

$$\begin{aligned} L = {} & [-\xi_1(1 - u(t))\beta_1 xy + \xi_2(1 - u(t))\beta_1 xy - w_1(t)u(t) \\ & - w_2(t)(1 - u(t)) + Ru^2] + \text{terms without u(t).} \end{aligned} \tag{47.9}$$

According to "Pontrygin's Minimum Principle", the unrestricted optimal control u^* satisfies $\frac{\partial L}{\partial u} = 0$ at $u = u^*$.

Therefore $u^*(t) = \frac{(\xi_2 - \xi_1)\beta_1 xy + w_1(t) - w_2(t)}{2R}$.

To determine the explicit expression for the optimal control without w_1 and w_2 including the boundary condition we consider the following three cases:

(i) For the set $\{t|0 < u^*(t) < 1\}$, we have $w_1(t) = w_2(t) = 0$, hence the optimal control is: $u^*(t) = \frac{(\xi_2 - \xi_1)\beta_1 xy}{2R}$.

(ii) For the set $\{t|u^*(t) = 1\}$ we have $w_1(t) = 0$, hence $u^*(t) = 1 = \frac{(\xi_2 - \xi_1)\beta_1 xy - w_2}{2R}$. Again since $w_2(t) > 0, \Rightarrow \frac{(\xi_2 - \xi_1)\beta_1 xy - w_2}{2R} \geq 1$.

(iii) For the set $\{t|u^*(t) = 0\}$, we have $w_2(t) = 0$. Hence the optimal control is $0 = u^*(t) = \frac{(\xi_2 - \xi_1)\beta_1 xy + w_1}{2R}$. Therefore $w_1(t) \geq 0 \Rightarrow \frac{(\xi_2 - \xi_1)\beta_1 xy}{2R} \leq 0$.

Combining these three cases, the optimal control is characterized as

$$u^*(t) = \max\left\{0, \min\left(1, \frac{(\xi_2 - \xi_1)\beta_1 xy}{2R}\right)\right\}. \tag{47.10}$$

If $\xi_1 > \xi_2$ for some t, then $u^*(t) \neq 1$. Thus $0 \leq u^*(t) \leq 1$ for such t means treatment should be administered. Hence we have a proposition.

Proposition *An optimal control u^* for system (47.6) maximizing the objective function (47.5) is characterized by (47.10). Thus we find the optimal control $u^*(t)$ for (47.6) computed with (47.8) together with (47.10). Here we have only treated the case $\delta \leq u(t) < 1$, $\delta > 0$, which means that chemotherapy never completely stopped viral replication.*

47.5.3 Uniqueness of the Optimality System

To prove the uniqueness of the solution of the optimal system we assume that the state system and adjoint system have finite upper bound depends on the condition $x(t) > T_m$.

Lemma *The function $u^*(t) = \max\{a, \min(b, s)\}$ is Lipschitz continuous in s, where $a < b$ are some fixed positive constant.*

Theorem *For t_f sufficiently small, bounded solution to the optimality system are unique.*

Proof is obvious.

From the above theorem, we can say that the solution of the system of such nonlinear boundary value problem is unique for a small time interval. For HIV patients the optimal control u^* gives an unique and optimal chemotherapy strategy.

Table 47.1 Variables and parameters used in the models (47.4) and (47.6)

Parameters	Definition	Default values assigned (day^{-1})
λ	Constant rate of production of $CD4^+T$	10.0 mm^{-3}[9]
p	Proliferation rate constant	0.03 [9]
T_m	Maximum proliferation of $CD4^+T$ cells	1500 mm^{-3} [9, 10]
d_1	Death rate of uninfected $CD4^+T$ cells	0.01 [9]
β_1	Rate of contact between x and y	0.002 mm^{-3}[13]
d_2	Death rate of virus producing cells	0.24 [9]
β_2	Killing rate of virus producing cells by CTL_d	0.001 mm^{-3}[9]
β_3	Killing rate of virus producing cells CTL_e	0.001 mm^{-3} [13]
s	Rate of simulation of CTL_d	0.2 [13, 14]
d_3	Death rate of CTL_d	0.02 [2]
c	Rate of simulation of CTL_e	0.2 [13, 14]
d_4	Death rate of CTL_e	0.02 [13, 14]

47.6 Numerical Solutions of the Model Equations

In this section we illustrate without control and with control model numerically (Table 47.1). In the numerical simulation we assume $x(0) = 100$, $y(t) = 50$, $z_d(t) = 2$, $z_e(t) = 5$.

For the numerical illustration of the optimal control problem (47.5) and (47.6) we assume $t_f = 1$, which can be used as an initial guess. We solve the optimality system by making the changes of the variable $\tau = t/t_f$ and transferring the interval [0, 1]. Here τ represents the step size which is used for better strategy with a line search method which will maximize the reduction of performance measure. We choose $t_f = 1 + \Delta t_f$ and initially $t_f = 1$. We also assume that $\Delta t_f = 0.1$ and our desired value of $t_f = 100$. The solution are displayed in Figs. 47.1, 47.2, 47.3, and 47.4. In Fig. 47.1 the number of uninfected T cell decreases rapidly with a short period and then reaches to its stable region. But if the control is used uninfected T cell population increases and reach to its adequate level. In this figure we also see that the number of infected T cell decreases but it does not extinct. While control is used the infected T cell moves to extinction. Here we also see that both the Lymphocyte cell population (Drug induced CTL and Effector CTL) decreases very fast for treated patients compare to uncontrolled problem.

Figure 47.2 shows that for different weight factors $R = 0.1$ and $R = 0.5$ numerical solution for optimal treatment strategy has been generated. The optimal system is solved using "The Steepest Descent Method". Here we also measure minimum cost $J = 0.8804$ for $R = 0.1$ and $J = 1.9612$ for $R = 0.5$. Thus for less value of R the optimal therapy $u^*(t)$ will achieve. In Fig. 47.3 we plotted two different cases using $R = 0.1$ and 0.5. All cases are during the treatment period. Here we see that uninfected T cell population decreases if the weight function R increases. Where as the infected T cell and both the lymphocyte cell population are decreases proportionate to R. In this figure we also observe that only uninfected T cell and drug induced CTL cell, for different value of R produce similar graph at

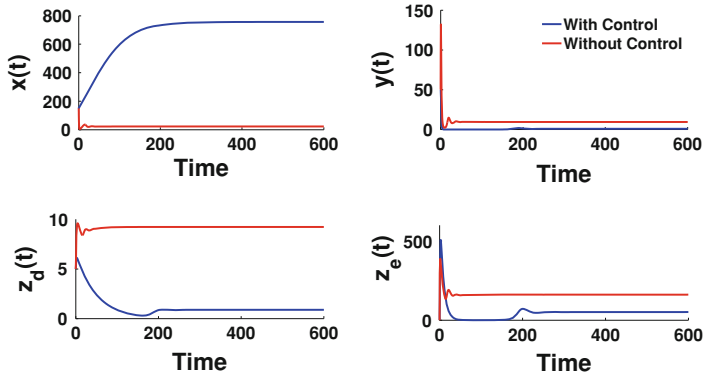
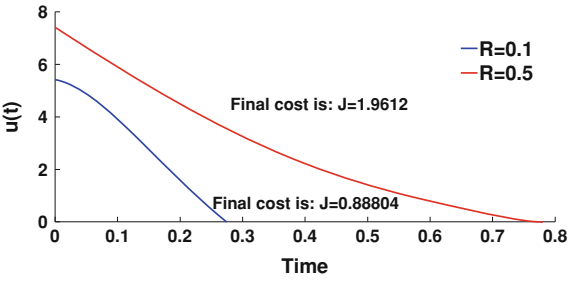


Fig. 47.1 Time series solution for the cases non-treated (without control) and treated (with control) keeping $u = 0.01$ and all other parameter as in Table 47.1

Fig. 47.2 Time series solution for the cases non-treated (without control) and treated (with control) keeping $u = 0.01$ and all other parameter as in Table 47.1



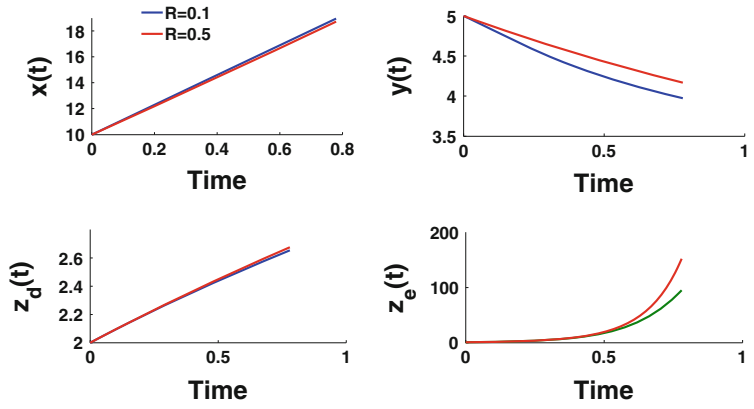


Fig. 47.3 The system behavior for optimal treatment when final time $t_f = 0.78$. Keeping all other parameter as in Table 47.1

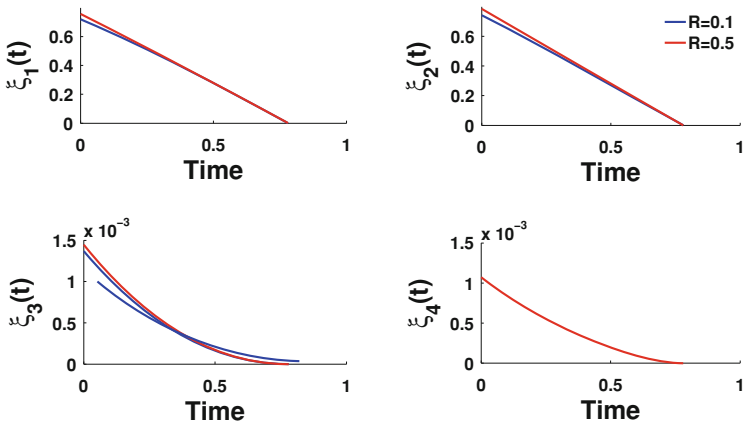


Fig. 47.4 The adjoint variable $\xi_1(t)$, $\xi_2(t)$, $\xi_3(t)$, $\xi_4(t)$ of the optimality system

initial stage. However after certain days of treatment the uninfected T cell population decreases down together with R but drug induced CTL population increases with R. In Fig. 47.4 we see that for different $R = 0.1$ and 0.5 there exist two different graph in all four cases, but as the time progresses the adjoint variables form similar graph irrespective of different R.

47.7 Discussion and Conclusion

In this paper we have formulated a mathematical model to study HIV dynamics within the human immune system including a control in the diseases transmission term. We have analyzed without the control model and control model, both by analytical techniques. After that we have done a comparative study both models by numerical techniques. In our analytical studies we have found out that the required conditions for different equilibrium points of the system. The system (47.4) has four equilibrium. The equilibrium E_0 is stable for *Proposition* 1, the equilibrium E_1 is stable for *Proposition* 2, the equilibrium E' exists for the *Proposition* 3 and lastly the interior equilibrium is asymptotically stable if it satisfy the *Proposition* 4. On the basis of existence condition we have calculated optimal control variable $u^*(t)$ for which the cost function will be minimized. Analytically we have seen that this optimal control variable is unique and derives the condition for which the system has its unique optimal control variable.

Thus we can conclude that immune system become active if long term drug therapy is introduced. But it is established that during uncontrolled treatment process the infected cell decreases yet it does not extinct. Long term drug therapy has a side effect. It is observed if drug therapy is introduced in an optimal control level the required result must appear that the infected cell will extinct and side by

side uninfected cell will also moves towards its steady state region. Here it should be emphasized from collecting data that the limit of the drug dose is near about 500 days. But from our numerical solution (Fig. 47.2), if it is used for less than 50 days the result for best treatment is to be appeared, but if it is introduced for more than 50 days the worst condition will appear inspite of a better one. Once infection is low, the immune response is not required at high levels and this is why it too drops off. We see that when the immune response are high, less medication is needed to control and regulate infection. Our optimal treatment is reduces the period of time while the immune response of the uninfected T cell takes over. So from mathematical calculation and numerical simulation we may infer that interruption of drug therapy is needed to allow to rebuild the immune system.

Acknowledgments Research is supported by the Government of India, Ministry of Science and Technology, Mathematical Science office, No. SR/S4/MS: 558/08.

References

1. Altes HK, Wodarz D, Jansen VAA (2002) The dual role of CD4T helper cells in the infection dynamics of HIV and their importance for vaccination. J Theor Biol 214:633–644
2. Bonhoeffer S, Coffin JM, Nowak MA (1997) Human immunodeficiency virus drug therapy and virus load. J Virol 71:3275–3278
3. Iwami S, Miura T, Nakaoka S, Takuchi Y (2009) Immune impairment in HIV infection: existence of risky and immunodeficiency thresholds. J Theor Biol 260:490–501
4. Kamien M, Schwartz NL (1991) Dynamic optimization, 2nd edn. North Holland, New York
5. Kim WH, Chung HB, Chung CC (2006) Optimal switching in structured treatment interruption for HIV therapy. Asian J Control 8(3):290–296
6. Kirshner D, Lenhart S, Serbin S (1997) Optimal control of the chemotherapy of HIV. J Math Biol 35:775–792
7. Kwon HD (2007) Optimal treatment strategies derived from a HIV model with drug-resistant mutants. Appl Math Comput 188:1193–1204
8. Lukas DL (1982) Differential equation classical to controlled mathematical in science and engineering. Academic Press, New York
9. Perelson AS, Krischner DE, De Boer R (1993) Dynamics of HIV infection of CD4 T cells. Math Biosci 114:81–125
10. Perelson AS, Neuman AU, Markowitz M, Leonard JM, Ho DD (1996) HIV 1 dynamics in vivo viron clearance rate infected cell life span and viral generation time. Science 271:1582–1586
11. Roy PK, Chatterjee AN (2010) T-cell proliferation in a mathematical model of CTL activity through HIV-1 infection. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, WCE 2010, 30 June–2 July, London, UK, pp 615–620
12. Skim H, Han SJ, Chung CC, Nan SW, Seo JH (2003) Optimal scheduling of drug treatment for HIV infection. Int J Control Autom Syst 1(3):282–288
13. Wodarz D, Nowak MA (1999) Specific therapy regimes could lead to long-term immunological control to HIV. Proc Natl Acad Sci USA 96(25):14464–14469
14. Wodarz D, May RM, Nowak MA (2000) The role of antigen-independent persistence of memory cytotoxic T lymphocytes. Int Immunol 12(A):467–477

Chapter 48
Design, Development and Validation of a Novel Mechanical Occlusion Device for Transcervical Sterilization

Muhammad Rehan, James Eugene Coleman and Abdul Ghani Olabi

Abstract The use of contemporary medical devices in the human body, such as dilation balloons, closure devices, stents, coils, stent-grafts, etc. are gaining more importance to preclude surgical incisions and general anaesthesia. An analogous procedure for permanent female sterilization is the transcervical approach that does not require either general anaesthesia or surgical incision and uses a normal body passage. Various contemporary technologies have improved the strategies for permanent female sterilization. However, current methods of transcervical sterilization are unable to provide an instant occlusion. This work presents the design, development and validation of a novel mechanical occlusion device (Rehan et al. 2010, Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July London, UK, vol I. pp 566–571), which achieve both instant and permanent female sterilization via a transcervical approach. The device is designed to provide an instant mechanical occlusion by deploying, under hysteroscopic visualization an implant into the intramural segment of the fallopian tube. The design of the device has been accomplished through computer aided design (CAD), finite element method (FEM) and experimental testing. Validation has been performed following a number of successful bench-top *in-air* and in vitro deployments on animal tissue and explanted human uteri. The efficacy of the device and the instant occlusion of the fallopian tubes were proved by hydraulic pressure testing of the implanted uteri using saline and methylene blue solution. Initial results suggest that the device

M. Rehan (✉) · A. G. Olabi
School of Mechanical and Manufacturing Engineering, Dublin City University, Glasnevin, D9, Ireland
e-mail: muhammad.rehan2@mail.dcu.ie

J. E. Coleman
Alta Science, Trinitas House, 2012, Orchards Avenue, Citywest, D24, Dublin, Ireland
e-mail: jecoleman@vasorum.ie

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_48,
© Springer Science+Business Media B.V. 2011

provides a safe, effective and instant method of permanent female sterilization. Further development work is ongoing in preparation for *first*-in vivo clinical trials.

48.1 Introduction

Surgical occlusion of the fallopian tubes is a widely used method of female sterilization because of its proven safety and effectiveness. The traditional surgical procedures are minilaparotomy and tubal ligation [1]. The advancements in these procedures have led to approaches such as electrocoagulation or clip or ring application to the tubes. However, such approaches to female sterilization are relatively high risk due to the requirement of general anesthesia with vascular damage, injury to the bowel, bladder, or uterus being potential complications. In addition, these procedures may be associated with postoperative pain [2].

The transcervical approach is an alternative to incisional procedures for interval tubal sterilization as it eliminates the requirement for general anesthesia and surgery. The most common methods of transcervical sterilization procedures depend on either destructive or mechanically occlusive approaches. Destructive methods have included chemical caustics, tissue adhesives, thermic induction and lasers [3, 4]. Destructive occlusion results in both a low success rate and morbidity [2, 3].

Contrary to the destructive methods of burning, freezing or fibrosing, the tubal ostia can be occluded by hysteroscopically applied mechanical devices. Such mechanical occlusion can be achieved either by placing a pre-formed plug or device in the uterotubal orifice or by formed-in situ methods. The technological developments in endoscopes, light transmission devices, optical resolution, catheters and tubal cannulation evolved some new technologies such as the Adiana and Essure devices [3]. However, both the Adiana [5, 6] and Essure [7–9] procedures rely on tissue in-growth from the surrounding tubal walls and are effective 3 months after device placement. This can be inconvenient for the patient, who has to use an alternate contraception during this time, which means an additional cost of contraception and a procedure to confirm tubal occlusion. Therefore, the requirement was to develop a transcervical approach that can provide an instant occlusion of the fallopian tube.

This work presents the design, development and verification of a novel mechanical occlusion device which achieves permanent female sterilization via the transcervical approach. Using a standard hysteroscope of 5-French (F) operating channel, the device deploy an implant [10] into the intramural section of the fallopian tube to provide an instant mechanical occlusion [11]. The device comprises of an implant, a guiding system and an actuator handle. The implant is made of biocompatible grade stainless steel (SS)316LVM and includes a guide tip at the distal end and a novel design of laser cut slots on the cylindrical body. These slots transform into two sets of wings that penetrate into the ostium and uterine muscle tissue entrapping the tissue and thereby plugging the entrance of the fallopian tube.

The ergonomically designed actuator controls the deployment and release of the implant at the target location by applying required forces in a specified sequence. The design of the device was achieved through CAD, FEA, prototyping and experimentations. The device was validated a number of times by successful deployments on the bench, in animal tissue and in explanted human uteri. During deployments in the latter, it was observed that the device provided both an instant and effective occlusion of the fallopian tube.

48.2 Materials and Methods

The device is designed, under hysteroscopic visualization to deploy an implant into the intramural segment of the fallopian tube to provide an instant mechanical occlusion [12].

48.2.1 Design and Development

The device consists of three major systems, an implant for occlusion of fallopian tube, a guide tube and wire combination for guidance of the implant through the cervix and an actuator handle to control the deployment and release of the implant, as shown in Fig. 48.1. The implant is attached at the flexible distal end of the guiding system. The proximal end of the guiding system is attached with the actuator handle. The occlusion system can be advanced through a 5-French (F) (1.67 mm internal diameter) operating channel of a standard hysteroscope. During insertion, the implant forms a low profile cylindrical shape and is advanced through the use of guiding system. After arriving at the target location within the human uterus, the required forces for the deployment and release of the implant are applied through the actuator in a specified sequence.

The implant The implant consists of a flexible guide tip at the distal end and a main cylindrical body housing an inner release system comprising of a core shaft and release tube as shown in Fig. 48.2. The implant main cylindrical body, with a length of 6.5 mm, an outer diameter (Ø) of 1.535 mm and a thickness of 0.1 mm, is made of annealed SS-316LVM. It features two sets of six slots at the distal and proximal segments. Post deployment, these slots determine the implant final shape by formation of two set of six wings. These wings serve to anchor the implant by protruding into the tubal ostium and entrapping the tissue of the intramural section to instantaneously occlude the fallopian tubes. The proximal end of the implant includes straight splines used to couple with the guide tube. Figure 48.3 depicts the comparison of the un-deployed and deployed implant. The guide tip is a Ø 0.5 mm, multi-filament ($7 \times 7 \times 7$) cable with a spherical ball shape at the distal end. The guide tip is designed to guide the implant through the uterus into the fallopian tube. Hence, a fine balance between column strength (for push-ability

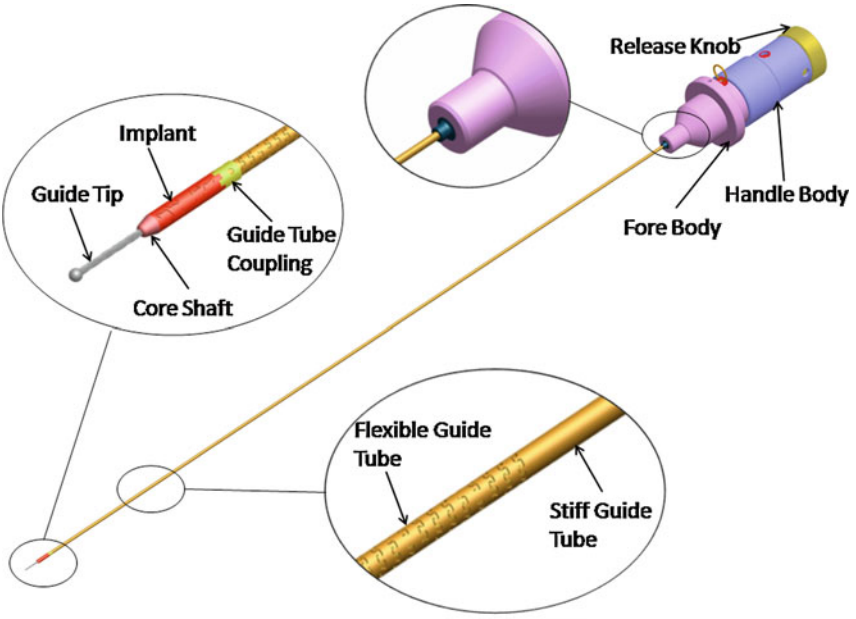


Fig. 48.1 Detailed view of the complete device

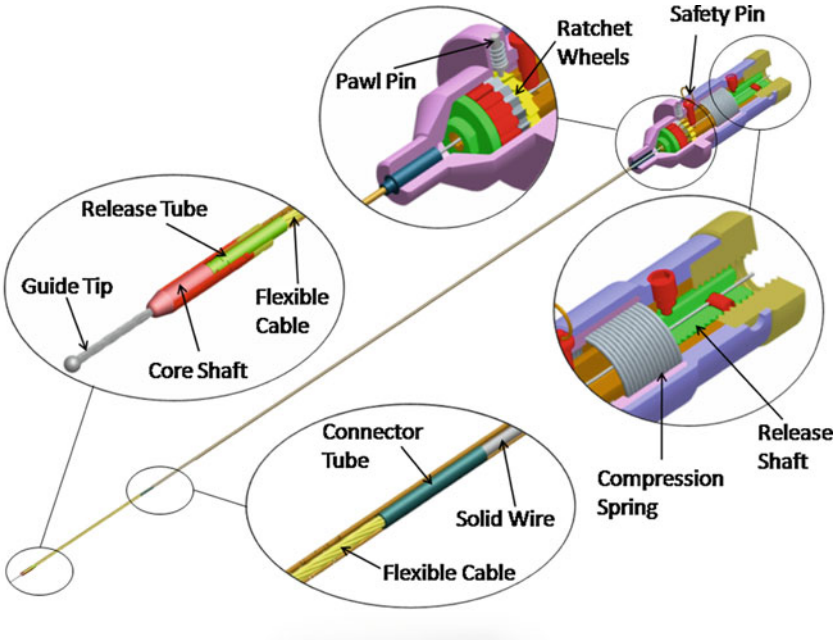


Fig. 48.2 Detailed sectional view of the complete device

and forward progression) and flexibility (to negotiate the curvatures of the uterus and fallopian tube) is required. The core shaft at the distal end of implant is a hardened SS-316LVM solid shaft, whose one end is conical and laser welded to the distal end of the implant and other end to the release tube. The release tube, with a length of 7 mm, an outer Ø of 1 mm and a thickness of 0.125 mm is made of hardened SS-316LVM tube and includes a pair of slots. This symmetric pair of laser cut slots forms an arc shape at both ends and a rectangular pattern in between. The gap in between this pair of slots forms a neck region which is designed to break at a specified load. Once the implant is deployed into the intramural segment of the fallopian tube, the weak link designed on the release tube is broken, releasing the implant from the guide system and consequently from delivery actuator.

Guide system The implant is delivered into the tubal ostium by the guiding system which includes an outer guide tube and an inner guide wire. The guide tube includes straight splines at the distal end, which are matched exactly with the implant splines. These matching straight splines are used to couple the guide tube with the implant as shown in Fig. 48.1. The inner guide wire is attached to the implant's release tube as shown in Fig. 48.2. In order to deal with the curvatures of the uterus and fallopian tube, the guide system needs to be flexible. On the other hand stiffness is required to transfer one-to-one torque to the implant. In order to acquire maximum torqability from the guiding system, a combination of flexibility and stiffness is designed into the guide system. As the device is delivered through the rigid channel of the hysteroscope and only a small distal portion of guide system emerges out beyond the hysteroscope channel. Therefore, only distal portions of the guide tube and wire were designed flexible. The guide tube, with a diameter (Ø) 1.3 mm and a thickness of 0.125 mm is made of SS-316LVM hardened tube. The flexibility at the distal end of the guide tube was achieved by the addition of segmented (inter-segment gap of 0.32 mm) chain of "dove tail" shaped helical slots with a pitch of 0.87 mm as shown in Fig. 48.1. These laser-cut slot shapes were designed to provide the required flexibility and torqability. In order to obtain flexibility at the distal end of the guide wire, a multi-filament cable was laser welded with a single rigid wire as shown in Fig. 48.2. Thus, the 360 mm long guide wire, comprises of a Ø 0.7 mm multi-filament (1 × 7) SS-316LVM cable with a length of 65 mm and a Ø 0.7 mm annealed SS-316LVM wire.

Actuator handle The proximal end of the guiding system is attached to the actuator handle. The material used for the components of the actuator is SS-316LVM. The actuator handle comprises a handle body of Ø 25 mm adapted to hold the actuator. At the distal end, a fore body is slidably and rotatably connected to the handle body. The fore body is also operatively connected to handle body through a ratchet mechanism in which the handle body incorporates a pair of ratchet wheels and the fore body includes a pawl pin as shown in Fig. 48.2. These ratchets are used to control the precise clockwise (CW) and counter-clockwise (CCW) movements by restricting any inadvertent reverse rotations. A compression spring is used in between the handle body and fore body to assist the movements in axial directions. The actuator also features a safety pin, which locks the actuator,

preventing accidental deployment during handling or transportation. This safety pin needs to be removed prior to deployment. The proximal end of the handle body includes a release mechanism to control the release of the implant as shown in Figs. 48.1 and 48.2. This release mechanism includes a threaded release shaft slidably connected to the handle body and a release knob rotatably connected to the handle body. The power screw mechanism in between the release knob and release shaft converts the applied torque through the release knob into a tensile force on the release shaft. The release torque was limited to a value that a human hand can apply with an index finger on a cylinder of Ø 25 mm. The guide tube connected to the fore body experiences compression and the guide wire connected to the release shaft experiences tension during clockwise rotation of the release knob. This results in breaking of the release tube from the specified location, releasing the implant.

48.2.2 Device Operation

The initial form of the implant is shown in Fig. 48.4. The device is designed to deploy and release in a five steps sequence.

Step 1 The CCW rotation of the fore body applies a 15.4 N-mm clockwise torque to the implant, which generates an out-of-plane displacement in the distal slots as shown in Fig. 48.5.

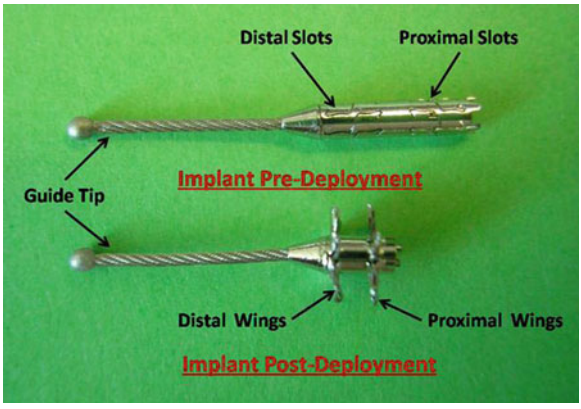


Fig. 48.3 Implant pre-deployment and post-deployment

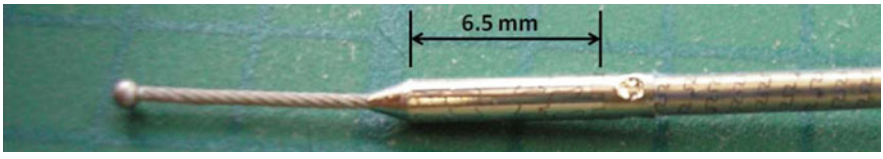


Fig. 48.4 Initial stage

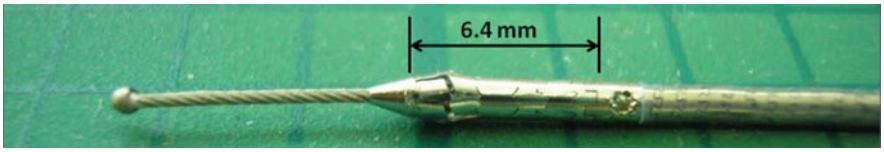


Fig. 48.5 Sequence step 1

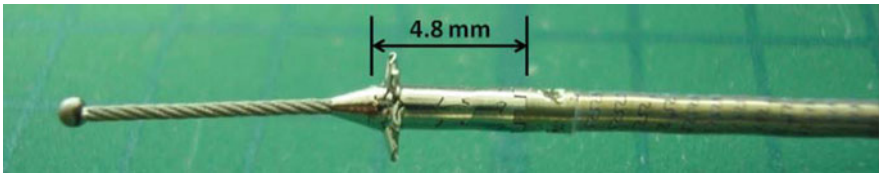


Fig. 48.6 Sequence step 2

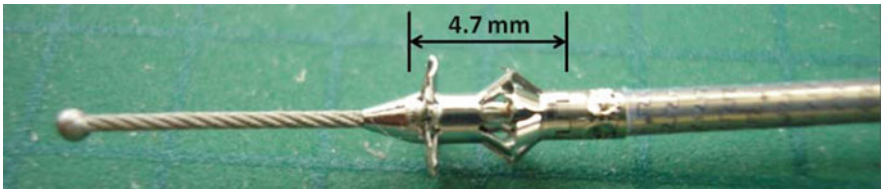


Fig. 48.7 Sequence step 3

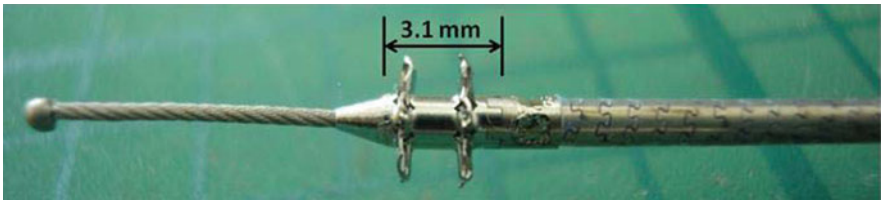


Fig. 48.8 Sequence step 4

Step 2 The compression spring applies a 30 N of axial compression force on the implant that plastically deforms the displaced slots into shape of the six distal wings as shown in Fig. 48.6.

Step 3 The CW rotation of the fore body applies a 16 N-mm counter-clockwise torque to the implant, which generates an out-of-plane displacement in the proximal slots as shown in Fig. 48.7.

Step 4 The compression spring applies a 25 N of axial compression force on the implant that plastically deforms the displaced proximal slots into shape of the six proximal wings as shown in Fig. 48.8.

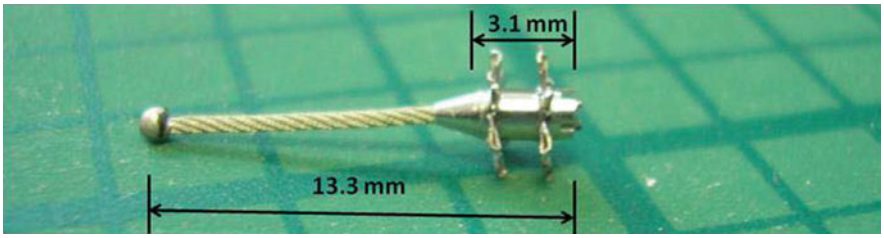


Fig. 48.9 Sequence step 5 (deployed implant)

Step 5 The CW rotation of release knob applies a 70 N of tensile force on the release tube allowing it to break at the designed location resulting in the release of the implant from actuator handle as shown in Fig. 48.9.

48.2.3 Validation

The device was evaluated a number of times ($n > 80$) in the laboratory. The evaluations include bench-top deployments of the implant *in-air*, *in-hysteroscopic* diagnostic model and in vitro. The bench-top *in-air* and in vitro testing were performed to assess efficacy of the device and validate the individual components including the implant, the delivery system and the actuator handle. In the in vitro study, the device was implanted in porcine tissue, arteries and fallopian tubes. Bench-top deployments involved the deployments of the device *in-air* to validate the functionality and mechanical behavior of the device.

The device was also validated using a hysteroscopic diagnostic model obtained from Limbs and Things, Bristol, UK. This model includes a universal base, a uterus clamp and three diagnostics uteri. These uteri covers all the conditions of the uterus including, normal anatomy, polyps and fibroids. Therefore, device was investigated in all three types of uterus conditions for its functionality and deliverability. In the testing, the diagnostic model was attached with a saline bottle to simulate an actual scenario. The hysteroscope was introduced to visualize the uterus anatomy as shown in Fig. 48.10. The device was then introduced into the uterus through the hysteroscope. During this testing, it was observed that the simulator provided the realistic simulation of device manoeuvres and the device was successfully delivered and deployed at the target location.

In vitro bench testing ($n = 30$) was carried out on both porcine tissue and fallopian tubes. These tests were performed to validate the deployment inside tissues against external loads, i.e. the loading exerted by tissue on the implant.

To evaluate the performance of the device in conditions very similar to in vivo implementation, in vitro experiments were conducted using explanted uteri. These uteri were removed at hysterectomy for various benign indications at the Midlands Regional Hospital, Mullingar, Ireland. Explanted uteri were chosen as the test

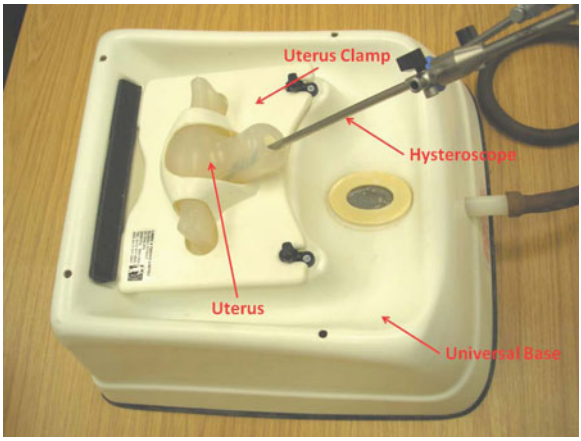


Fig. 48.10 Device testing in hysteroscopic diagnostic model

model as this is most representative model of the in vivo situation. These studies ($n = 11$) were performed to validate the functionality, deliverability and effectiveness of the device for instant closure of human fallopian tubes. The device was delivered and deployed bilaterally into the tubal ostia. A small caliber hysteroscope with a 5-F operating channel was used to deliver the implant into the ostium tissue. The uterus was distended with normal saline and spurt of saline from the fallopian tubes confirmed the un-obstruction. The hysteroscope was introduced under direct vision into the uterus and both tubal opening were observed as shown in Fig. 48.11a. After positioning, the device was guided through the hysteroscope as shown in Fig. 48.11b. On approaching the tubal ostium, the implant was positioned in the intramural segment of the fallopian tube until the straight splines at the implant distal end became invisible as shown in Fig. 48.11c. After optimal placement, the implant was deployed and released from the delivery actuator as shown in Fig. 48.11d. The delivery system was withdrawn and the procedure was repeated on the contra-lateral tube. After successful deployment of implants in both tubes, a hydraulic pressure test of the uterus was performed to verify occlusion of the fallopian tubes. In this test, saline and methylene blue solution was introduced into the uterus at a pressure of 300 mmHg. The pressure was held for 5 min to ensure the blockage of the fallopian tubes. Finally, the ostium and tubes were dissected to examine the placement and deployment of the implant in the intramural section of the ostium. The implant along with some tissue was extracted to further examine the wing shape, deployed implant and tissue entrapped.

48.3 Results and Discussion

The device for transcervical sterilization presented in this work has various advantages over tubal sterilization: avoiding general anesthesis, no incision in the body, a clinical procedure and decreased cost. Its main advantage over other

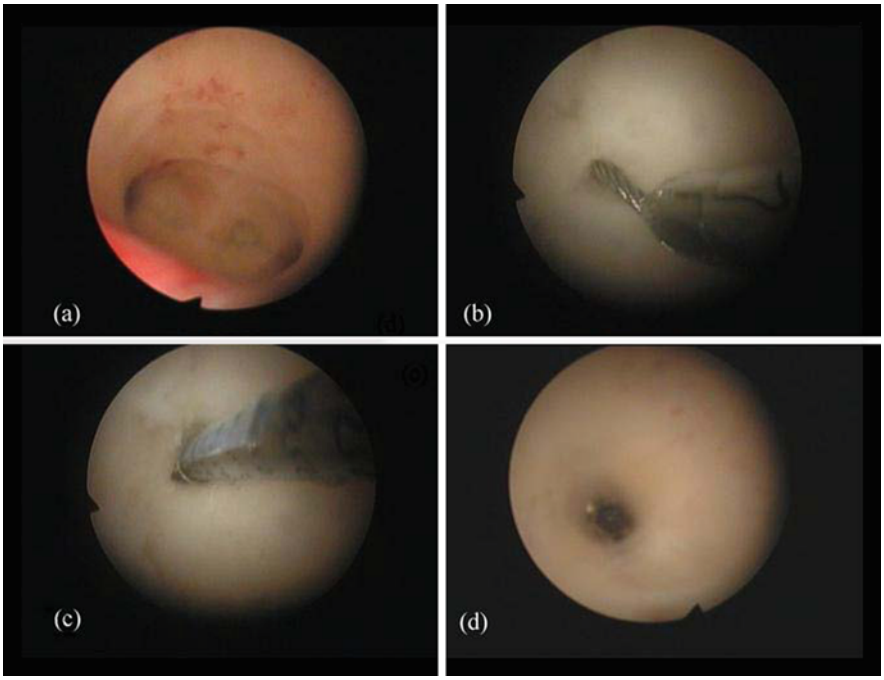


Fig. 48.11 Hysteroscopic views of Xplant studies: **a** tubal ostia, **b** implant guide tip ingoing left tubal opening, **c** implant optimally placed at left tubal ostium, **d** deployed Implant at left tubal ostium

transcervical procedures is the instant mechanical occlusion to effect female sterilization. The design of the device, 3D modeling and FEA simulations were performed using CAD and FEA software. The device was validated by experimentation and testing. The implant was fabricated using laser cutting machine (LPL Stent Cutter). The fabricated implant was deployed under microscope and its mechanical behavior was studied. The wing profile of the deployed implants was measured using video inspection probe and its profile was compared with the designed profile. The standard error of mean of the difference of experimental and designed profile was 0.003543 and the maximum percent error was 3.129%.

The forces required to deploy implants were validated experimentally on the test bench. A 16.0 N-mm moment produced a 0.353 mm out-of-plane displacement in implant slots. A comparable 15.4 N-mm was measured experimentally using a torque meter when same amount of radial expansion (from Ø1.535 to 2.241 mm) in implant slots was achieved. In order to plastically deform these expanded implant slots, a force of 25.4 N was measured from tensile testing which is comparable to a force of 25 N obtained from the design. In-house in vitro and mechanical bench testing validated the mechanical behavior and functional aspects of the implant.

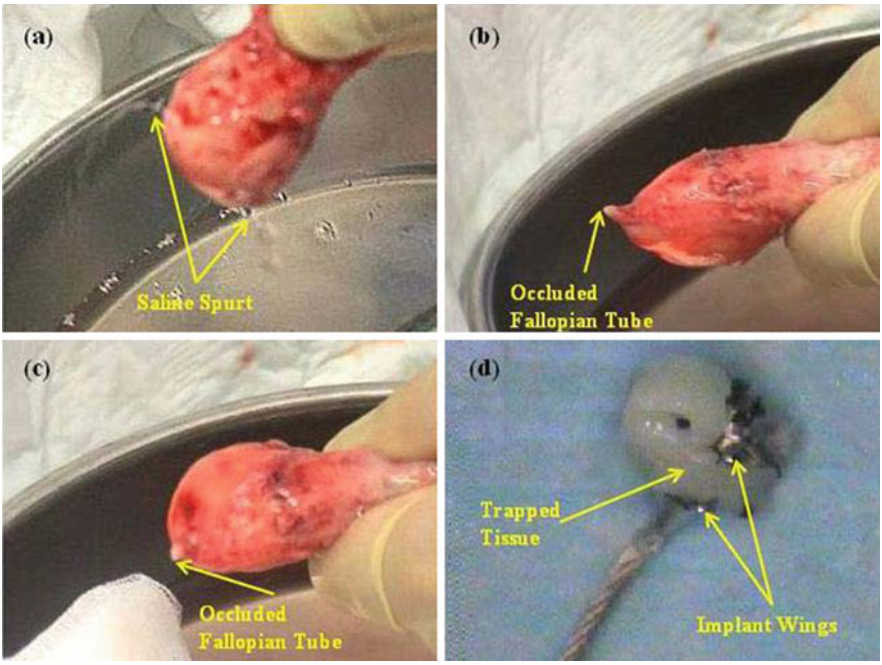


Fig. 48.12 Uterus in Xplant studies: **a** un-obstructed fallopian tubes pre-deployment, **b** occluded fallopian tubes post-deployment, **c** hydraulic pressure testing post-deployment, **d** deployed implant in dissected

In vitro deployments ($n = 11$) of the implant into human explanted uteri were performed. Figure 48.12a shows the un-obstructed fallopian tubes before deployment. The occluded fallopian tube is shown in Fig. 48.12b. As expected, the device had successfully occluded the fallopian tubes in all uteri. Immediately after the bilateral deployment of device in the uteri, the hydraulic pressure tests using saline water and methylene blue solution at pressure of 300 mmHg were performed. It is apparent from Fig. 48.12c that there was no leakage during these pressure testing. These uteri were dissected after pressure testing and the implant along with some tissue was examined under microscope to further investigate the implant wings shape, deployed implant and tissue trapped as shown in Fig. 48.12d. It was observed in dissection that there was no indication of methylene blue after distal wings of the implant. This demonstrates the capability of the device to achieve instant occlusion of the fallopian tubes even at a pressure of 300 mmHg.

The statistical methods used to analyse and report the results of explants studies were the statistical summaries of results and tabulation of the data. Occlusion of each fallopian tube was treated as an individual event representing 0 or 1 for "no" or "yes" occlusion, respectively. Bilateral deployments of the implant were attempted in 11 explant studies. Successful instant occlusion was achieved in 22 out of 22 (100%) fallopian tubes. Since there were zero failures among 11

explanted uteri, statistical significance was not established because of zero numerator problems. However, these effective in vitro tests are important in the development of the device as they provide a clear indication of the device's capabilities and weaknesses prior to commencing in vivo verification and validation activities. Data generated from investigations in explanted uteri verified the efficacy of the implant as well as minimizing the risk to patients participating in in vivo clinical studies. Further development work is underway in preparation for *first*-in vivo clinical trials. As an overall conclusion, this implant is effective, consistent and feasible for hysteroscopic occlusion of fallopian tube.

48.4 Conclusion and Future Work

This novel device is a new prospect for transcervical sterilization, which uses inherent body channel and provides an immediate occlusion of the fallopian tubes. This device is minimally invasive, safe, effective, easy to use, consistent, reproducible and feasible for the hysteroscopic occlusion of fallopian tubes.

The device presented has been successfully validated the instant and effective occlusion of fallopian tubes for permanent female sterilization. The following future work is recommended:

1. About 180 million [1] couples per annum are relying on female sterilization to avoid further pregnancies. However, the sterilization procedures of permanent nature have discouraged some couples from choosing these methods. Therefore, much work is still required to make this device an effective method of reversible hysteroscopic female sterilization.
2. The concept and design of the complete device could be used for other medical applications.
3. The actuator concept could be implemented in various medical and aerospace applications, where rotational and translations displacements are required.

Acknowledgments The authors thank Vasorum/Alta Science team for their support. The authors thank Dr. Michael Gannon from the Midlands Regional Hospital, Mullingar, Ireland for the arrangement and performing in vitro studies in human uteri.

References

1. Thurkow AL (2005) Hysteroscopic sterilization: away with the laparoscope? Int Congress Ser 1279:184–188
2. Magos A, Chapman L (2004) Hysteroscopic tubal sterilization. Obstet Gynecol Clin North Am 31:31705–31719
3. Abbott J (2005) Transcervical sterilization, Best Pract Res Clin Obstet Gynaecol 19:743–756

4. Brumsted JR, Shirk G, Soderling MJ, Reed T (1991) Attempted transcervical occlusion of the fallopian tube with the Nd:YAG laser. Obstet Gynecol 77:327–328
5. Carr-Brendel VE, Stewart DR, Harrington DC, Leal JGG, Vancaillie T (2001) A new transcervical sterilization procedure: results of a pilot implant study in humans. Obstet Gynecol 97:8–5
6. Carr-Brendel VE, Stewart DR, Harrington DC, Dhaka VK, Breining PM, Vancaillie T (2001) A new transcervical sterilization procedure—6-month preclinical results. Obstet Gynecol 97:15–16
7. McSwain H, Brodie MF (2006) Fallopian tube occlusion, an alternative to tubal ligation. Tech Vasc Interv Radiol 9:24–29
8. Ubeda A, Labastida R, Dexeus S (2004) Essure®: a new device for hysteroscopic tubal sterilization in an outpatient setting. Fertil Steril 82:196–199
9. Cooper JM, Carignan CS, Cher D, Kerin JF (2003) Microinsert nonincisional hysteroscopic sterilization. Obstet Gynecol 102:59–67
10. Rehan M, Coleman JE, Olabi AG (2010) Novel implant for transcervical sterilization. J Biosci Bioeng 110(2):242–249
11. Wimer B, Dong RG, Welcome DE, Warren C, McDowell TW (2009) Development of a new dynamometer for measuring grip strength applied on a cylindrical handle. Med Eng Phys 31:695–704
12. Rehan M, Coleman JE, Olabi AG (2010) Novel mechanical occlusion device for transcervical sterilization. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, 30 June–2 July, London, UK vol I. pp 566–571

Chapter 49
Investigation of Cell Adhesion, Contraction and Physical Restructuring on Shear Sensitive Liquid Crystals

Chin Fhong Soon, Mansour Youseffi, Nick Blagden and Morgan Denyer

Abstract In this study, the nature of cell attachment and contraction on the surface of a shear sensitive cholesteryl ester liquid crystal (LC) was examined. This interaction has the potential to be utilized as a novel cell force transducing assay. Preliminary studies indicated that cells cultured on the LC induce deformations in the underlying LC layer. This study aimed at determining if those deformations resulted from the weight of the cell or from forces generated within the cell being transmitted to the LC surface by focal adhesions (FA). In order to study this cell-surface relationship, the forces generated within the cell by the actin cytoskeleton were inhibited by treatment with 30 μM cytochalasin-B and cell surface attachment via integrins was broken by treatment of cells with, 0.25% Trypsin–EDTA. In the study of the morphology changes of cells and their interfacial interactions within the LC were investigated using fluorescence staining of the actin cytoskeleton and Widefield Surface Plasmon Resonance (WSPR) microscopy. Both cytochalasin-B and trypsin treatments caused deformations in the shear sensitive LC surface to decrease and disappear. This indicates that the deformations in the LC were induced by forces generated in the actin cytoskeleton being transmitted to the LC surface via FA. Fluorescent staining of the actin cytoskeleton

C. F. Soon (✉) · N. Blagden · M. Denyer
School of Life Sciences, University of Bradford, Bradford, BD7 1DP, UK
e-mail: scfhong@bradford.ac.uk

N. Blagden
e-mail: n.blagden@bradford.ac.uk

M. Denyer
e-mail: m.denyer@bradford.ac.uk

M. Youseffi
School of Engineering Design and Technology-Medical Engineering,
University of Bradford, Bradford, BD7 1DP, UK
e-mail: m.youseffi@bradford.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*,
Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_49,
© Springer Science+Business Media B.V. 2011

and immunofluorescent vinculin staining indicated that cells cultured on the soft LC substrate developed a diffuse actin cytoskeleton and vinculin staining revealed FA around the periphery of the cells. These finding were confirmed by WSPR microscopy which indicated that cell surface attachments formed around the periphery of cells grown on the liquid crystals.

49.1 Introduction

Many of the biological activities of cells; such as cytokinesis, migration, morphogenesis, etc., are dependant on how cells adhere and contract [1, 2]. Attachment to the cell–extracellular matrix is mediated by the integrins that bind to the extracellular matrix (ECM) proteins (e.g., collagen, fibronectin, matrigel, etc.). Strengthened adhesion can be achieved via stress applied by the actin cytoskeleton via the integrin-focal adhesions linkage to the ECM. Previous silicon membrane based force measurement methods lacked the flexibility to discriminate cell forces exerted at high spatial resolution [2]. Furthermore, the requirement to functionalize the surface with ECM proteins has raised the concern about specific responses of cells to the type of protein used or the underlying stiffness of the substrate [3]. To mitigate these concerns, shear sensitive cholesteryl ester liquid crystals was proposed to be a novel cell force transducing assay. By employing this liquid crystal as a biosensor, incorporating elasticity and flexibility, was found to be capable of responding to culture with a human keratinocytes cell lines (HaCaTs) by forming deformations [4]. The attachment of cells was mediated without the need of surface functionalization with exogenous ECM proteins [5]. Other than being biocompatible [6], the cholesteryl ester liquid crystals when being immersed in the cell culture media are able to transform into a lyotropic system. Lyotropic liquid crystals are the basic structures of many living systems such as the lipid bilayers (LB) that make up the cell membrane [7]. These amphiphilic molecules such as the cholesteryl esters found on the LB can form a stable water insoluble monolayer [8]. Unlike the polymer based silicon membrane which is made up of molecules that are highly cross-liked, liquid crystals (LC) consist of compactly arranged molecules (or mesogens) which can function as a highly flexible cell culture substrate capable of sensing localized shear perturbations due to the cell's actin filament contraction or relaxation. Consequently such an investigation on the cells adhesion and physical restructuring of the surface of the LC, requires manipulation of the forces generated within a cell. To this end, cytochalasin-B and EDTA–Trypsin were used to disrupt the cytoskeleton and cell attachment, respectively. Cytochalasin-B inhibits cell division by inhibiting formation of contractile actin filaments [9, 10], and has been found to de-polymerize actin filament [11] and subsequently prohibit keratinocyte migration [12]. Trypsinization was also used to digest integrins on the surface of the cell membrane, hence preventing cells from attaching to any surface. To gain more insights into the actin cytoskeleton

organization, the restructuring of the cells actin filaments and focal adhesions (e.g., vinculin) on a soft liquid crystal substrate was studied by use of a fluorescence staining technique. The interfacial interaction of cells/liquid crystals were interrogated by the Widefield Surface Plasmon Resonance (WSPR) microscopy which has been shown to provide high contrast and high resolution imaging at the interface [13].

49.2 Aims and Objectives

The long term aim of our research is to develop a liquid crystal based cell force transducer which is able to detect the physical re-structuring of HaCaT cells and to measure forces generated by single cells. The aim of the present study was to investigate the nature of cells adhesion, contraction and physical restructuring on the surface of biocompatible liquid crystals. We investigated the liquid crystals sensitivity to the changes of adhesion and contraction. The physical responses of the cells on the soft liquid crystals were investigated and compared with cell adhesion on a glass surface.

49.3 Materials and Methods

49.3.1 Liquid Crystal Coating and Cell Culture

The preparation of the shear sensitive cholesteryl ester liquid crystals and coating were as described previously [5]. After the cover slip was coated, suspensions of HaCaT cell were acquired using a standard culture procedure [5]. The cells were plated onto liquid crystal coated glass cover slips in petri dishes. Then, the culture was added with RPMI-1640 media (Sigma–Aldrich, UK), which is supplemented with L–Glutamine (2 mM), penicillin (100 units/ml), streptomycin (100 mg/ml), Fungizone (2.5 mg/l) and 10% of Fetal Calf Serum. Exposure of the cholesteryl ester liquid crystals to the RPMI-1640 culture media resulted in the formation of the Lyotropic liquid crystals [5].

49.3.2 Treatment with Cytochalasin-B and EDTA-Trypsin

HaCaTs were cultured on the liquid crystal substrate in three separate culture petri dishes at a cell density of 500 cells/cm^2 and incubated at 37°C for 24 h [6]. On the following day, two of the samples were treated with Cytochalasin-B (Sigma–Aldrich) was diluted in 0.042% (v/v) ethanol in distilled water (35 mg/ml) and

same dilution of ethanol as control, respectively. The media in the last sample was replaced with 3 ml of 0.25% of EDTA–crude trypsin.

For the cytochalasin-B and EDTA–trypsin treatment, the petri dishes were placed on a hot stage maintained at 37°C for 5-min, and time-lapse images were captured using a phase contrast microscope for an hour. These experiments were repeated three times.

49.3.3 Fluorescence Staining of F-Actin and Vinculin

Cells were cultured on a plain glass cover slip and a glass cover slip coated with liquid crystals at a density of 1.3×10^3 cells/cm^2. At sub-confluency of cells, all the plain glass cover slips were removed from the petri dishes and washed twice with Hanks Balanced Salt Solution (HBSS, Sigma-Aldrich). Then, the cells were fixed with 1% formaldehyde in HBSS for 6 min, rinsed twice with HBSS and permeabilized with 0.1% Triton X-100 for 3 min. For F-actin staining after washing, the cells were incubated for 45 min with 1 μg/ml of Fluorescence isothiocyanate (FITC) labeled Phalloidin solution (Sigma-Aldrich, UK) in HBSS followed by another three washes. To stain the nuclei, 0.1 μg/ml of DAPI dihydrochloride solution (Sigma-Aldrich, UK) in HBSS was applied to the cells for 15 min.

For vinculin staining, sub-confluent cells were washed, fixed and permeabilized as described before. Followed by three washes in HBSS, the cells were bathed with 2% of bovine serum albumin (BSA from Sigma-Aldrich, UK) for 30 min. After blocking, cells were washed three times in HBSS and incubated in 50 μl of anti-human vinculin antibody (1:40) in 1% BSA for 24 h in a humidity chamber at 5°C. The procedures were repeated except that the primary antibody was omitted for the negative controls. After the incubation for 24 h, the substrates were washed three times with HBSS, blotted and incubated with goat anti mouse Immunoglobulin (IgG) secondary antibody labeled with Alexor Fluor 488 (5 μg/ml diluted in 1%BSA) for an hour. After staining, the substrate was subjected to three 5 min washes in HBSS before mounting onto the glass slides. These experiments were repeated three times.

49.3.4 Widefield Surface Plasmon Resonance (WSPR) Microscopy

Two round glass substrates with a diameter of 22 mm and coated with 50 nm thick gold layer were used in this study. The coating of cholesteryl ester liquid crystals on the gold substrate were prepared using the procedures described previously [4]. HaCaT cells at a density of 1.0×10^3 cells/cm^2 were plated in the two petri dishes, each containing one gold cover slip coated with and without liquid crystal coating. Subsequently, the two petri dishes were incubated at 37°C for 24 h.

After incubation, the cells were washed in HBSS twice and fixed in 1% formaldehyde for 6 min. The fixation was followed by a wash in HBSS and the cells were dehydrated using 100% ethanol serial dilution method for 5 min in each dilution. When the cells were dehydrated, each gold cover slip was transferred to the sample holder just above the objective lens of the WSPR microscope for imaging. The objective lens used was a Zeiss Plan Fluar of 1.45 Numerical Aperture (NA) with 100x magnification and with oil immersion contact. In the WSPR system used, an incident light from a Helium–Neon (HeNe) 633 nm laser source was used to excite the surface plasmons at the gold layer by p-polarized light at an excitation angle of 46°. The light wave interacts with the interface medium of different reflective index and generates various levels of reflectivity in the surface plasmon. The reflected light containing information about the interfacial interactions was captured by a charged coupled device (CCD) camera. A detail description of the optical system used has been reported elsewhere [6].

49.3.5 Statistical Analysis

Cell area and the length of discrete vinculin stained region were determined using the WSPR images and staining micrographs via ImageJ software, respectively. The required data (N) was expressed as mean ± standard deviation (SD). The data was analyzed by using independent-samples t test available in the Statistical Package for Social Sciences (SPSS) software.

49.4 Results and Discussion

49.4.1 Effects of Cytochalasin-B and EDTA–Trypsin to Cells Attached to Liquid Crystals

Cells generated stress marks in the liquid crystal coated substrate. These stress marks may be caused by the mass of the cells, electrostatic charges displacing the mesogens of the liquid crystal or by forces generated from the actin cytoskeletons within the cells to the liquid crystal surface via the focal adhesions. Therefore, the objective of this experiment is to determine the source of the forces inducing stress marks on the LC surface. After 24 h culture on the liquid crystals, the adherent HaCaTs induced some intense deformation lines (dark field/bright field stratification) on the surface of the LC in the three cultures prepared. For the control and two test culture experiments, three cells of similar size were chosen. Under the phase contrast microscopy, HaCaTs attached to LC were not fully spread but ruffles could be observed at the cell periphery before any treatment (Fig. 49.1, 0 min). When ethanol diluted in the distilled water was added to the control

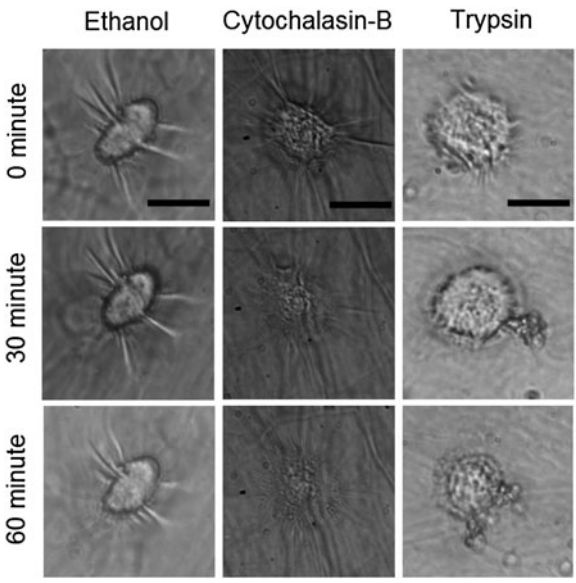


Fig. 49.1 Application of single dose 0.042% Ethanol (control), 30 μM Cytochalasin-B in 0.042% Ethanol, 0.25% EDTA–Trypsin solution on HaCaTs adhered on the surface of the lyotropic liquid crystals in 60 min treatment (scale bar: 25 μm)

(containing culture media and the solvent) in 60 min, no major changes of cell activity and deformation lines of the liquid crystal were observed (Fig. 49.1, Ethanol). Probably, this is due to the very low concentration of ethanol.

Addition of cytochalasin-B to the cultures caused remarkable changes in both cell morphology and liquid crystal's deformation lines distribution when compared with the control. Time lapsed images taken every 30 min (Fig. 49.1, cytochalasin-B) showed the effects to HaCaTs and the deformation lines correlated with the cells response. After 15 min of treatment, HaCaT cells had not changed their morphology but the deformation lines had started to shrink (Image not shown). After 30 min, cell attachment area seemed to increase and the ruffles on the cell body were greatly reduced. This is an indication of a release on the intracellular forces. By the end of the 60 min treatment, the cells remained attached, relaxed but they had acquired flattened cell morphology. The intense deformation lines in the liquid crystal changed into fine lines (Fig. 49.1, Cytochalasin-B).

Following the de-polymerization of actin filaments, the stress generated in the cells was reduced and this was implied by the reduction of deformation lines. From this experimental result, it is clear that cytochalasin-B has inhibited the function of actin filament in supporting HaCaTs contraction. In this case, Cytochalasin-B did not detach cells from the surface indicating that the actin polymerization response in generating the force within the cells were not associated with anchoring the cells to the surface, instead, they were mediated by the focal adhesions [14]. This attachment to the liquid crystals was probably maintained by the cell surface receptors and focal adhesions. Focal adhesions are a group of specific macromolecules assemblies such as vinculin, paxillin or talin connected to a pair of $\alpha\beta$ integrins that directly bind to the ECM [15]. This is also where the mechanical

force from actin cytoskeleton is transmitted to the extracellular matrix [15]. In contrast, we could not find this type of transducing effects when cells attached to the surface of a petri dish (stiff surface) [6] when treated with cytochalasin-B. Comparatively, other soft substrate force measurement method such as PDMS, polyacrylamide and collagen sheet [2, 15, 16] could not provide such a flexibility and localized deformations as presented by liquid crystals because of the rigidity because of the rigidity of these polymers.

Experiments involving trypsinization were used to confirm intracellular forces were transmitted to the liquid crystals via focal adhesions. The trypsinisation effect on HaCaTs culture over liquid crystals was very different from previous result where a total loss of liquid crystal deformation line was observed (Fig. 49.1, Trypsin). Crude EDTA–Trypsin cleaves the adhesion proteins that bound cells to the substrate, caused the periphery of the cell edges to detach and cell spreading to reduce (Fig. 49.1, Trypsin). After 60 min of trypsinization, the cells acquired a more rounded morphology and lifted off the surface of liquid crystals (Fig. 49.1, Trypsin). This indicated that the actin filaments had restructured to sustain a more rounded morphology. Despite the absence of the exogenous ECM proteins, the surface integrin receptors were still able to attach to the surface of the liquid crystals. We are not clear how the surface integrin receptors are bound to the liquid crystals but for this study we have assumed that is probably mediated by the self derived endogenous ECM proteins [16, 17].

49.4.2 Transverse Strain and Effects of Compression

Apart from causing the cell morphology changes, we discovered that cytochalasin-B experiment not only induced the relaxation of the cells but also showed that the growing direction of the deformation lines is directly correlated with the cells "pinching" the surface [18]. A group of contracted cells with elongation shapes was treated with 30 μM cytochalasin-B over a 60 min period on a hot stage set at 37°C. Before treatment with cytochalasin-B, contracted cells indented the surface of the LC with localized deformation lines growing in outward directions, perpendicular to the cell contraction direction as seen in Fig. 49.2. The reaction of the LC deformation lines could be explained by Newtons' third law of action and reaction. The stronger the compression, the more intense is the deformation line as seen using a microscope.

After treatment with cytochalasin-B for 30 min, the deformation lines of the liquid crystals induced by the cells were very much shortened in an inward direction perpendicular to the cell relaxation direction (Fig. 49.2c). From Fig. 49.2a and c, the relaxation direction shown is opposite to the cell contraction direction (solid line and dotted line arrows). The relaxation of HaCaTs was verified by an increase in cells length (123 μm) compared with the original cells length (102 μm). Clearly, the displacement of the longitudinal deformation line was due to the transverse strain at a finite region on the LC surface (indicated by

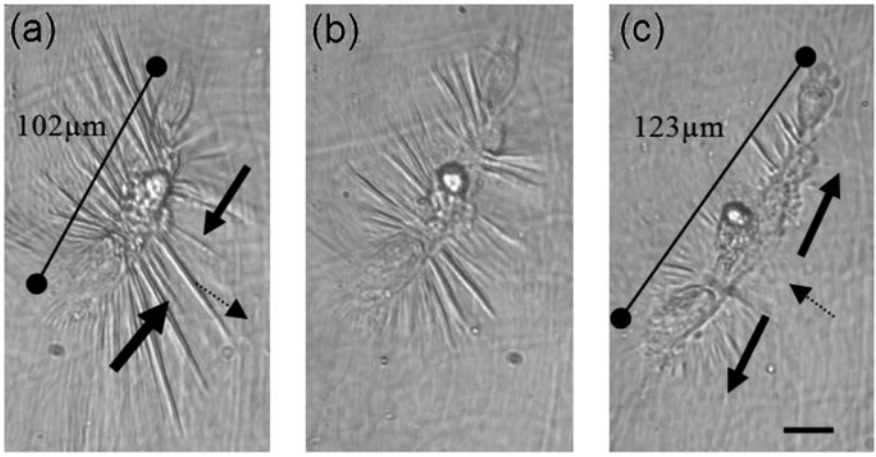


Fig. 49.2 HaCaTs on LLC in response to 30 μM cytochalasin-B and direction of CELC deformation lines shrinkage at **a** 0 min, **b** 30 min and **c** 60 min. *Solid line arrows* show the direction of the cell contraction and relaxation after treated with cytochalasin-B. *Dotted line arrows* show the repelling directions of the deformation line (scale bar = 25 μm)

solid line arrows in Fig. 49.2a). The transverse strain can be explained by the mechanism of the focal adhesions compressing the stress sites that are driven by the simultaneous active sliding and binding of actin filaments via the actin cross-linking proteins (e.g., myosin-II and α-actinin) [19]. Hence, the result shows that the longitudinal deformation line is a secondary effect of compression.

49.4.3 Structure of F-Actin and Vinculins

The F-actin and vinculin expression for cells on the plain glass substrate and the liquid crystal coated substrate were distinctly different. Cells plated on the hard surface of the glass substrate projected linear and striated arrays of stress fibers span the entire cells body (Fig. 49.3a). The enhanced prominent actin fibers was reflected by a higher expression of elongated and scattered vinculin staining in those cells attached to the glass surface (Fig. 49.3c). In comparison, more diffused actin filaments (Fig. 49.3b) accompanied vinculins in small punctuations (Fig. 49.3d) that are arranged at the periphery of cells membrane were observed for adherent cells on the liquid crystal substrate. The length of the vinculin stained was significantly different ($P < 0.001$, $n = 525$) with 1.4 ± 0.44 and 1.05 ± 0.42 μm for cells cultured on the plain glass when compared with the cells cultured on liquid crystal substrate (Fig. 49.4a), respectively. In the context of the area covered by individual cells, the cells on the glass covered at significantly greater area (2496 ± 608 μm^2) than the cells on the liquid crystal coated substrates (759 ± 169 μm^2, $P < 0.001$, $n = 258$) (Fig. 49.4b). The distinctly

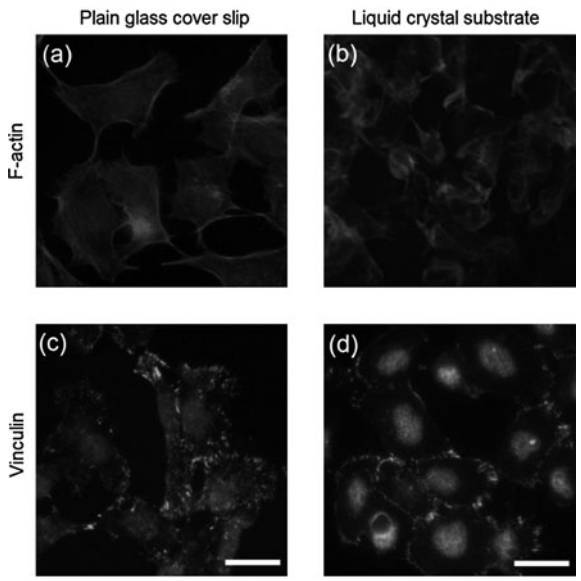


Fig. 49.3 Fluorescence micrographs of the staining against **a, b** F-actin and **c, d** vinculin (scale bar = 25 μm)

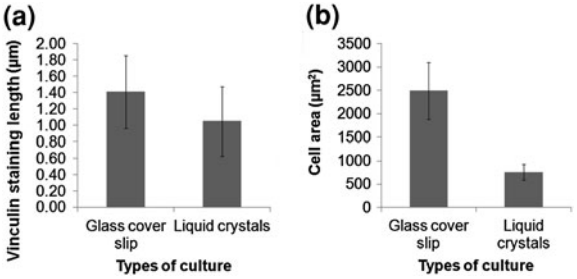


Fig. 49.4 Comparison of the **a** vinculin size and **b** cell covered area for adherent HaCaT cells on a plain glass cover slip and liquid crystal coated substrate. The area and length of vinculin are expressed in mean ± SD and the level of significant at $P < 0.001$, $N = 258$ (cell surface area) and $N = 525$ (vinculin)

different morphology of the cells on the two different types of culture substrates must have a strong correlation with the stiffness of liquid crystals underlying the cells. There is evidence that cells have an active mechano-sensing mechanism that monitor the stiffness of a substrate leading to a re-organization of the cytoskeleton [20, 21].

The dynamic behavior of cells in adapting to the stiffness of the liquid crystal substrate has induced significantly smaller vinculin staining regions or dots of vinculin that were diffusely arranged at the edges of the cell membrane. This occurred especially with a higher intensity at the cell–cell border. It is possible that such characteristics are very similar to the physical structures of keratinocytes

found in the epidermis layer [22]. These results indicate a positive interaction between HaCaT cells and the liquid crystals because the liquid crystals act as a biomaterial mimicking a cell's native environment [23]. In addition, this view is strengthened by the observation that HaCaT cells grown on the liquid crystals continued to differentiate and proliferate after several days of cultures. If cells change their physical structure and interactions on substrate of different stiffness, then, a cell force transducer developed on a material with a higher or lower elastic modulus than that of the in vivo system would measure different range of exertion forces [20, 24] and may even be taken as pathological condition by the cells [24]. Hence, designing a cell mechano-transducer using biomaterials with elastic modulus closer to that of the tissue would literally provide relevant mechanical signal to the cells. This is a point that requires stringent consideration in the development of cell force transducer systems.

Overall the evidence and findings of this study supports the perception that the rounded morphology of HaCaT cells on the liquid crystal substrate was associated with a decrease in the assembly of the F-actin and regular structure of focal adhesions as seen in the immunostaining. The two elements (employment of actin filament and adhesion plaques) are inter-related and are believed to occur in a closed-loop dependency [14, 25]. While attaching to a soft substrate, a cell disassembles the focal adhesion proteins via the Rho-GTPase pathways and the disassembly of adhesion proteins, in turn, promotes de-phosphorylation of the actin cytoskeleton driven by Myosin II [19]. Due to the down regulation of the cell focal adhesions and cytoskeletal depolymerization on the soft liquid crystal substrate, cells decrease their physical contact with the liquid crystals and increase their contacts with the neighboring cells (Fig. 49.3b, d).

49.4.4 Interfacial Interactions of Cells and Liquid Crystals

Fluorescence staining was used to show the internal structure and re-organization of the cells in responding to the contact surface and WSPR images revealed the interfacial stress as a result of cell restructuring. For the WSPR microscopy, HaCaTs were widely spread on the plain gold cover slip. A wide band with irregular distribution of punctuated patterns was formed around the nucleus located at the centre of the cell (middle ring) and this indicated a greater contact area at the cell/glass interface (Fig. 49.5a). This is in good correlation with the vinculin staining (Fig. 49.3c), and the irregular patterns on the wide band (Fig. 49.5a) seemed to be related to the stress points by scattered focal adhesions on the plain glass substrate [13]. Contrarily, the adherent cell on the liquid crystals was found to be "less spread" and in restricted contact areas with the surfaces. Clearly defined narrow and yet a lower number of concentric rings were seen around the nucleus or periphery of the cell on the LC (Fig. 49.5b). Consistently, the concentric rings and the regular narrow bands at the cell boundary were associated with the weight of the nucleus and uniform focal adhesion arrangements

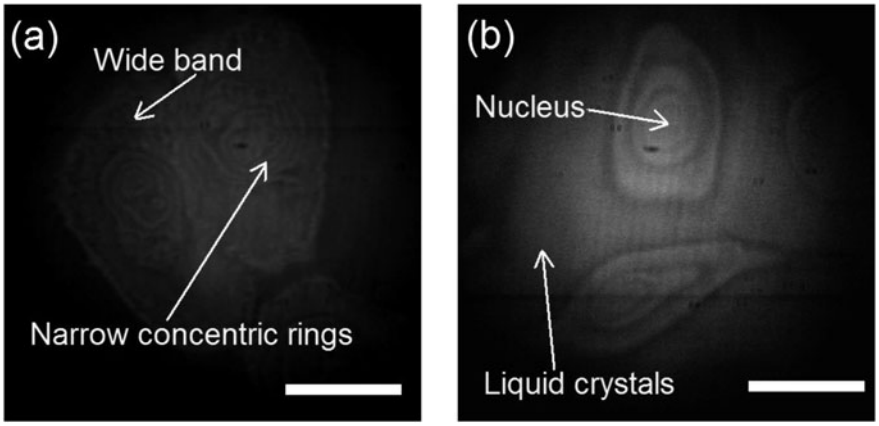


Fig. 49.5 WSPR images showing interfacial interactions of HaCaTs cultured on **a** a plain glass cover slip, and **b** liquid crystals (scale bar = 20 μm)

as indicated in Fig. 49.3d, respectively. This suggested that the stress applied by the focal adhesions induced by the actin filaments was concentrated around the edges of the cells membrane in a regular and narrow contact area. The higher contrast regions shown in Fig. 49.5b are due to the presence of the liquid crystals interfacing with the cells. The result also showed that the attachment mediated by the focal adhesion on the liquid crystal substrate was sufficient to prevent cells being removed by washing during the preparation of the substrate before being imaged using the immunofluorescence or WSPR microscope. Nonetheless, this experiment again showed that the physical interaction with the liquid crystals had influenced the cells morphology. The recognition of the external mechanical signal by the cells may have overtaken the biochemical signal in regulating the cell surface area couplings [23].

49.5 Conclusions

The results of this study clearly indicate that a cholesteryl ester based lyotropic liquid crystals not only facilitate cell adhesion via cell adhesion proteins but can also induce cytoskeletal re-organization. Furthermore, the results indicate the deformation lines in the shear sensitive liquid crystals formed around the cells are caused by contraction forces generated within the cell via actin filaments and these forces are transmitted to the soft liquid crystal surface via focal adhesions.

Acknowledgments Also, we wish to thank Dr. Steve Britland, Dr. Peter Twigg and Dr. Samira Batista for their helpful discussions about this work. Appreciation to Dr. Samira Batista Lobo, Dr. ShuGang Liu and Sali Khagani for their technical support.

References

1. Alberts B, Johnson A, Lewis J, Raff M (2000) Molecular biology of the cell. Garland Science, New York
2. Beningo A, Wang YL (2002) Flexible substrata for the detection of cellular traction forces. TRENDS Cell Biol 12(2):79–84
3. Engler A, Bacakova L, Newman C, Hategan A, Griffin M, Discher D (2002) Substrate compliance versus ligand density in cell on gel responses. Biophys J 86:617–628
4. Soon CF, Youseffi M, Blagden N, Lobo SB, Javid FA, Denyer MCT (2009) Interaction of cells with elastic cholesteryl liquid crystal. IFBME Proc 25/X:9–12
5. Soon CF, Youseffi M, Blagden N, Denyer MCT (2010) Effects of an enzyme, depolymerization and polymerization drugs to cells adhesion and contraction on lyotropic liquid crystals. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2010, WCE 2010, vol. 1, 30 June–2 July, 2010, London, UK, pp 556–561
6. Soon CF, Youseffi M, Blagden N, Lobo SB, Javid FA, Denyer MCT (2009) Characterization and biocompatibility study of nematic and cholesteryl liquid crystals. In: Lecture notes in engineering and computer science: Proceedings of world congress on engineering, WCE 2009, vol 2, 1–2 July, 2009, London, UK, pp 1872–1875
7. Small DM (1977) Liquid crystals in living and dying systems. J Colloid Interface Sci 58(3): 581–602
8. Hata Y, John J, Insull W (1974) Cholesteryl ester-rich inclusions from human aortic fatty streak and fibrous plaque lesions of atherosclerosis. Am J Pathol 75(3):423–456
9. Theodoropoulos PA (1974) Cytochalasin B may shorten actin filaments by a mechanism independent of barbed end capping. Biochem Pharmacol 47(10):1875–1881
10. Smith GF, Rider MAC, Janet F (1967) Action of cytochalasin B on cultured human lymphocytes. Nature 216:1134–1135
11. Stournaras C, Köhler R, Rössle M, Zentel R (1996) Altered actin polymerization dynamics in various malignant cell types: evidence for differential sensitivity to cytochalasin B. Biochem Pharmacol 2:1339–1346
12. Morioka S, Lazarus GS, Baird JL, Jensen P (1987) Migrating keratinocytes express urokinase-type plasminogen activator. J Investig Dermatol 88:418–423
13. Jamil MMA, Denyer MCT, Youseffi M, Britland S, Liu ST, See CW, Somekh MG, Zhang J (2008) Imaging of the cell surface interface using objective coupled wide field surface plasmon microscopy. J Struct Biol 164:75–80
14. Geiger B, Bershadsky A (2001) Assembly and mechanosensory function of focal contacts. Curr Opin Cell Biol 13(5):584–592
15. Bershadsky A, Balaban NQ, Geiger B (2003) Adhesion-dependent cell mechanosensitivity. Ann Rev Cell Dev Biol 19:677–695
16. Kirfel G, Herzog H (2004) Migration of epidermal keratinocytes: mechanisms, regulation and biological significance. Protoplasma 223:67–68
17. O' Toole EA (2001) Extracellular matrix and keratinocyte migration. Clin Exp Dermatol 26:525–530
18. Oliver T, Dembo M, Jacobson K (1995) Traction forces in locomoting cells. Cell Motil Cytoskeleton 31:225–240
19. Pelligrin S, Mellor H (2007) Actin stress fibres. J Cells Sci 120:3491–3499
20. Yeung T, Georges PC, Flanagan LA, Marg B, Ortiz M, Funaki M, Zahir N, Ming W, Weaver V, Janmey PA (2005) Effects of substrate stiffness on cell morphology, cytoskeletal structure, and adhesion. Cell Motil Cytoskeleton 60:24–34
21. Discher DE, Janmey P, Wang YL (2005) Tissue cells feel and respond to the stiffness of their substrate. Science 310:1139–1143
22. Koegel H, Tobel LV, Schafer M, Alberti S, Kremmer E, Mauch C, Hohl D, Wang XJ, Beer HD, Bloch W, Nordheim A, Werner S (2009) Loss of serum response factor in keratinocytes results in hyperproliferative skin disease in mice. J Clin Invest 119(4):899–910

23. Lo CM, Wang HB, Dembo M, Wang YL (2000) Cell movement is guided by the rigidity of the substrate. Biophys J 79:144–152
24. Engler A, Griffin MA, Sen S, Bonnemann CG, Sweeney HL, Discher DE (2004) Myotubes differentiate optimally on substrates with tissue-like stiffness pathological implications for soft or stiff microenvironments. J Cell Biol 166(6):877–887
25. Geiger B, Bershadsky A (2002) Exploring the neighborhood: adhesion-coupled cell mechanosensors. Cell Press J 110(2):139–142

Chapter 50
On the Current Densities for the Electrical Impedance Equation

Marco Pedro Ramirez Tachiquin, Jose de Jesus Gutierrez Cortes, Victor Daniel Sanchez Nava and Edgar Bernal Flores

Abstract Based upon the modern Pseudoanalytic Function Theory, we discuss the analytic structure of the electrical current densities for the three-dimensional generalized Ohm's law, obtained bias the so-called Taylor series in formal powers when the conductivity is a scalar separable-variables function.

50.1 Introduction

This chapter is mainly based upon the material published in the previous work [14], but it has been complemented with results that will allow us to pose the elements of what can be considered a new Theory for the Electrical Impedance Tomography problem, based mainly on the Applied Pseudoanalytic Theory [8], and some of its generalizations for Quaternionic Analysis. We intend to show most of basic ideas as well as the patches that might be followed in order to have a deeper comprehension of the conductivity phenomena for the static case. We shall start remarking that the Electrical Impedance Tomography problem was posed by Calderon [4], basically a boundary value problem for the partial differential equation

$$\nabla \cdot (\sigma \nabla u) = 0, \tag{50.1}$$

where σ is the conductivity function, and u is the electric potential. For this problem, we should approach σ inside a domain Ω with a boundary Γ, given the values of u at such boundary.

M. P. R. Tachiquin (✉) · J. de Jesus Gutierrez Cortes · V. D. S. Nava · E. B. Flores
Facultad de Ingenieria de la Universidad La Salle, B. Franklin 47, 06140 México C.P., Mexico
e-mail: marco.ramirez@lasallistas.org.mx

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_50,
© Springer Science+Business Media B.V. 2011

As it is well known, this problem is specially important for medical imaging, because it represents an auxiliary noninvasive technique for the diagnosis of several diseases, as such characterized by the presence of a certain class of tumors, being also useful for delicate medical observation processes, as it is the neonatal lung monitoring [2].

Yet, by more than 20 years after Calderon formally posed the Electrical Impedance Tomography problem, the mathematical complexity of (50.1) restricted the use of this technique because the images obtained by means of it were quite deficient when compared to such provided by, for instance, the Positron Emission Tomography, or the Magnetic Resonance Imaging.

But in 2006, by relating (50.1) with a complex differential equation usually know as *Vekua equation* [16], Astala and Päivärinta [1] gave a positive answer for the two-dimensional Electrical Impedance Tomography problem. And in 2007, Kravchenko [10] obtained what it could be considered the first general solution of (50.1) in analytic form, employing elements of the Pseudoanalytic Function Theory [3], for a certain class of σ.

Following the interesting path opened by these two discoverings, in 2009 it was possible to pose the general solution for the two-dimensional case of (50.1) in terms of Taylor series in formal powers, when σ is a separable-variables function.

By considering a quaternionic generalization of the Bers generating pair for complex-valued functions [11], we will study the structure of the general solution for the quaternionic three-dimensional Electrical Impedance Equation, and bias a quaternionic generalization of the Beltrami equation in complex analysis, we will review how to obtain an infinite set of analytic solutions for (50.1), when the conductivity σ is a separable-variables function depending upon three spacial variables.

Our discussion will finish by considering one particular case of the conductivity σ for which it is possible to obtain in exact form a set of *formal powers*, and in consequence, the electrical current patches inside the bounded domain Ω.

50.2 Elements of Quaternionic Analysis and Pseudoanalytic Function Theory

The algebra of real quaternions (see e.g. [9]) will be denoted by $\mathbb{H}(\mathbb{R})$. The elements q belonging to $\mathbb{H}(\mathbb{R})$ have the form $q = q_0 + \overrightarrow{q}$, where q_0 is usually referred as the *scalar part* of q, and $\overrightarrow{q} = \sum_{k=1}^{3} \mathbf{i}_k q_k$ is the *vectorial part* of q. The elements $q_k \in \mathbb{R}$; $k = \overline{0,3}$ whereas $\mathbf{i}_1, \mathbf{i}_2$ and $\mathbf{i}_3$ are the quaternionic units possessing the properties

$$\mathbf{i}_1\mathbf{i}_2 = -\mathbf{i}_2\mathbf{i}_1 = \mathbf{i}_3, \quad \mathbf{i}_2\mathbf{i}_3 = -\mathbf{i}_3\mathbf{i}_2 = \mathbf{i}_1, \quad \mathbf{i}_3\mathbf{i}_1 = -\mathbf{i}_1\mathbf{i}_3 = \mathbf{i}_2, \quad \mathbf{i}_1^2 = \mathbf{i}_2^2 = \mathbf{i}_3^2 = -1.$$

The subset of purely vectorial quaternions $q = \overrightarrow{q}$ can be conveniently identified with the set of real three-dimensional vectors $\mathbb{R}^3$. This is, every three-dimensional vector $\overrightarrow{E} = (E_1, E_2, E_3)$ can be associated with a purely vectorial quaternion $\overrightarrow{E} = E_1\mathbf{i}_1 + E_2\mathbf{i}_2 + E_3\mathbf{i}_3$. It is evident that the correspondence is one to one.

By this isomorphism, the multiplication between two vectorial quaternions $\overrightarrow{q} = \sum_{k=1}^{3} \mathbf{i}_k q_k$ and $\overrightarrow{p} = \sum_{k=1}^{3} \mathbf{i}_k p_k$ can be written employing standard vectorial notations:

$$\overrightarrow{q}\,\overrightarrow{p} = -\overrightarrow{q} \cdot \overrightarrow{p} + \overrightarrow{q} \times \overrightarrow{p}, \tag{50.2}$$

where $\overrightarrow{q} \cdot \overrightarrow{p}$ is the classical Cartesian scalar product, and $\overrightarrow{q} \times \overrightarrow{p}$ represents the vectorial product in the quaternionic sense

$$\overrightarrow{q} \times \overrightarrow{p} = (q_2p_3 - q_3p_2)\mathbf{i}_1 + (q_3p_1 - q_1p_3)\mathbf{i}_2 + (q_1p_2 - q_2p_1)\mathbf{i}_3.$$

The reader can easily notice that, in general, $\overrightarrow{q}\,\overrightarrow{p} \neq \overrightarrow{p}\,\overrightarrow{q}$. Because of this, we will use the notation

$$M^{\overrightarrow{p}}\,\overrightarrow{q} = \overrightarrow{q}\,\overrightarrow{p} \tag{50.3}$$

to indicate the *multiplication by the right-hand side* of the quaternion $\overrightarrow{q}$ by the quaternion $\overrightarrow{p}$.

On the set of at least once differentiable functions, we can define the Moisil–Theodoresco differential operator $\mathbf{D} = \mathbf{i}_1\partial_1 + \mathbf{i}_2\partial_2 + \mathbf{i}_3\partial_3$, where $\partial_k = \frac{\partial}{\partial x_k}$. According to the classical vectorial notations, when applying $\mathbf{D}$ to a quaternion $\overrightarrow{q}$ we obtain

$$\mathbf{D}\overrightarrow{q} = -\nabla \cdot \overrightarrow{q} + \nabla \times \overrightarrow{q}. \tag{50.4}$$

Notice also that any scalar function (let us say σ) can be considered a purely scalar quaternion, and there for

$$\mathbf{D}\sigma = \nabla\sigma. \tag{50.5}$$

Let us now introduce some concepts of the Pseudoanalytic Function Theory posed first in [3]. Let F and G be two complex-valued functions satisfying

$$\mathrm{Im}(\overline{F}G) > 0, \tag{50.6}$$

where $\overline{F}$ denotes the complex conjugation of $F : \overline{F} = \mathrm{Re}F - i\,\mathrm{Im}\,F$, and i denotes the standard complex unit $i^2 = -1$. Thus any complex-valued function W can be represented by means of the linear combination of these two functions:

$$W = \phi F + \psi G,$$

where ϕ and ψ are real-valued functions. When a pair of functions (F, G) satisfies the condition (50.6), they are called a *Bers generating pair*.

This idea was used by Lipman Bers to develop the Pseudoanalytic Function Theory [3], which in recent years have become of great importance for the Electrical Impedance Tomography Theory, as the authors intend to pose in the following pages.

Bers introduced the (F, G)-*derivative* of a complex-valued function W (now also called in his honor *derivative in the sense of Bers*) as follows:

$$\partial_{(F,G)} W = (\partial_z \phi) F + (\partial_z \psi) G, \tag{50.7}$$

where $\partial_z = \partial_x - i\partial_y$, and it will exist if and only if the following equation is fulfilled

$$(\partial_{\overline{z}} \phi) F + (\partial_{\overline{z}} \psi) G = 0, \tag{50.8}$$

where $\partial_{\overline{z}} = \partial_x + i\partial_y$.

In order to express (50.7) and (50.8) in terms of W, let us introduce the notations

$$\begin{aligned} A_{(F,G)} &= \frac{\overline{F}\partial_z G - \overline{G}\partial_z F}{F\overline{G} - \overline{F}G}, \quad a_{(F,G)} = -\frac{\overline{F}\partial_{\overline{z}} G - \overline{G}\partial_{\overline{z}} F}{F\overline{G} - \overline{F}G}, \\ B_{(F,G)} &= \frac{F\partial_z G - G\partial_z F}{F\overline{G} - \overline{F}G}, \quad b_{(F,G)} = -\frac{G\partial_{\overline{z}} F - F\partial_{\overline{z}} G}{F\overline{G} - \overline{F}G}. \end{aligned} \tag{50.9}$$

These expressions are called the *characteristic coefficients* of the generating pair (F, G). Applying the new notations, Eq. 50.7 becomes

$$\partial_{(F,G)} W = \partial_z W - A_{(F,G)} W - B_{(F,G)} \overline{W}, \tag{50.10}$$

and (50.8) turns into

$$\partial_{\overline{z}} W - a_{(F,G)} W - b_{(F,G)} \overline{W} = 0. \tag{50.11}$$

Equation 50.11 is known as the *Vekua equation*, because it was deeply studied by Ilia Vekua [16]. A complex-valued function W that fulfills the equation (50.11) will be called (F, G)-pseudoanalytic.

The following statements were originally presented in [3], and they have been slightly adapted for this work.

The complex-valued functions that constitute the generating pair (F, G) are (F, G)-pseudoanalytic, and their (F, G) -derivatives are $\partial_{(F,G)} F = \partial_{(F,G)} G = 0$.

Definition 50.1 Let (F_0, G_0) and (F_1, G_1) be two generating pairs, and let their characteristic coefficients satisfy

$$a_{(F,G)} = a_{(F_1,G_1)} \quad \text{and} \quad B_{(F,G)} = -b_{(F_1,G_1)}. \tag{50.12}$$

Thus, the pair (F_1, G_1) will be called the *successor pair* of (F_0, G_0), as well (F_0, G_0) will be named the *predecessor pair* of (F_1, G_1).

Theorem 50.2 *Let the W be* (F_0, G_0)*-pseudoanalytic, and let* (F_1, G_1) *be a successor pair of* (F_0, G_0)*. Then the* (F_0, G_0)*-derivative of W will be* (F_1, G_1)*-pseudoanalytic*.

Bers also posed the (F, G)*-integral* of a complex-valued function W, as it will be explained in the following paragraphs.

Definition 50.3 Let (F_0, G_0) be a generating pair. Its *adjoint pair* will be denoted by (F_0^*, G_0^*), *and will be defined by the formulas*

$$F_0^* = -\frac{2\overline{F}_0}{F_0\overline{G}_0 - \overline{F}_0 G_0}, \quad G_0^* = \frac{2\overline{G}_0}{F_0\overline{G}_0 - \overline{F}_0 G_0}.$$

Definition 50.4 The (F_0, G_0)-integral of a complex-valued function W is defined as

$$\int_{z_0}^{z_1} W\, d_{(F_0,G_0)}z = F_0(z_1)Re\int_{z_0}^{z_1} G_0^* W\, dz + G_0(z_1)Re\int_{z_0}^{z_1} F_0^* W\, dz. \tag{50.13}$$

where $z_0,\ z_1$ are fixed points in the complex plane, and it will exist if and only if

$$Re\, G_0^* W\, dz + i\, Re\, F_0^* W\, dz = 0.$$

Indeed, if $W = \phi F_0 + \psi G_0$ is an (F_0, G_0) -pseudoanalytic function, then

$$\int_{z_0}^{z} \partial_{(F,G)} W\, d_{(F,G)}z = W(z) - \phi(z_0)F(z) - \psi(z_0)G(z), \tag{50.14}$$

and since $\partial_{(F,G)}F = \partial_{(F,G)}G = 0$, the integral expression (50.14) represents the antiderivative in the sense of Bers of $\partial_{(F,G)}W$.

Theorem 50.5 *Let the generating pair* (F_0, G_0) *be a predecessor pair of* (F_1, G_1)*. A complex-valued function W will be* (F_1, G_1)*-pseudoanalytic if and only if it is* (F_0, G_0)*-integrable*.

Definition 50.6 Let $\{(F_m, G_m)\}$, $m = 0, \pm1, \pm2, \pm3, \ldots$ be a sequence of generating pairs. If every (F_{m+1}, G_{m+1}) is a successor pair of (F_m, G_m), we will call $\{(F_m, G_m)\}$ a *generating sequence*. Specifically, if the generating pair $(F_0, G_0) = (F, G)$ we say that (F, G) is *embedded in the generating sequence* $\{(F_m, G_m)\}$.

Definition 50.7 The generating pairs (F_m, G_m) and (G'_m, F'_m) are called equivalent if their characteristic coefficients (50.9) are the same.

Definition 50.8 A generating sequence$\{(F_m, G_m)\}$ is called periodic with period $\beta > 0$ if the generating pairs (F_m, G_m) and $(F_{m+\beta}, G_{m+\beta})$ are equivalent.

Definition 50.9 Let W be an (F, G)-pseudoanalytic function, and suppose that $\{(F_m, G_m)\}$ is a generating sequence in which (F, G) is embedded. Then we will be able to express the higher derivatives in the sense of Bers of W as

$$W^{[0]} = W; \quad W^{[m+1]} = \partial_{(F_m,G_m)} W^{[m]}; \quad m = 0, 1, 2, \ldots$$

Definition 50.10 The complex valued-function $Z_m^{(0)}(a_0, z_0; z)$ will be called *formal power with center* z_0, *coefficient* a_m *and exponent* 0. This function will depend upon a complex variable z, and it will be defined as the linear combination of the elements of the generating pair (F_m, G_m)

$$Z_m^{(0)}(a_0, z_0; z) = \lambda F_m + \mu G_m,$$

with a pair of real constant coefficients λ and μ such that

$$\lambda F_m(z_0) + \mu G_m(z_0) = a_0.$$

The *formal powers* with exponents $n = 1, 2, 3, \ldots$ will be defined according to the recursive formulas

$$Z_m^{(n+1)}(a_{n+1}, z_0; z) = (n+1) \int_{z_0}^{z} Z_{m+1}^{(n)}(a_{n+1}, z_0; \xi)\, d_{(F_m,G_m)}\xi,$$

where the integral operator is an integral in the sense of Bers, described in (50.13).

Theorem 50.11 *The formal powers posses the following properties*:

(1) $Z_m^{(n)}(a_n, z_0; z)$ *is an* (F_m, G_m)*-pseudoanalytic function.*
(2) *If* a_n' *and* a_n'' *are real constants, then*

$$Z_m^{(n)}(a_n' + i a_n'', z_0; z) = a_n' Z_m^{(n)}(1, z_0; z) + a_n'' Z_m^{(n)}(i, z_0; z).$$

(3) *The following relations hold*

$$\partial_{(F_m,G_m)} Z_m^{(n+1)}(a_n, z_0; z) = (n+1) Z_{m+1}^{(n)}(a_n, z_0; z).$$

(4) *And finally*,

$$Z_m^{(n)}(a_n, z_0; z) \sim a_n (z - z_0)^n \quad \textit{when } z \to z_0.$$

Every complex-valued function W, solution of (50.11), can be expanded in Taylor series in formal powers

$$W = \sum_{n=0}^{\infty} Z^{(n)}(a_n, z_0; z), \tag{50.15}$$

where the absence of the subindex m means that all formal powers correspond to the same generating pair. In other words, the expansion (50.15) is an analytic representation of the general solution of (50.11).

50.3 Analytic Solutions for the Electrical Impedance Equation

Consider the electrical impedance equation (50.1) with a conductivity function σ of the form

$$\sigma(x_1, x_2, x_3) = \alpha(x_1)\beta(x_2)\gamma(x_3). \tag{50.16}$$

Here α, β, and γ are real-valued functions at least once differentiable. Let us introduce the notations (see e.g. [15])

$$\overrightarrow{\mathcal{E}} = \sqrt{\sigma}\nabla u, \quad \overrightarrow{\sigma} = \frac{\nabla\sqrt{\sigma}}{\sqrt{\sigma}}, \tag{50.17}$$

where the gradient is applied in the sense of (50.5). Thus the Eq. 50.1 will turn into

$$(\mathbf{D} + M^{\overrightarrow{\sigma}})\overrightarrow{\mathcal{E}} = 0, \tag{50.18}$$

where $\mathbf{D}$ is the Moisil–Theodoresco differential operator (50.4) and $M^{\overrightarrow{\sigma}}$ is the operator of multiplication by the right-hand side (50.3). Since σ is a separable-variables function, it will be useful to introduce the auxiliary notations

$$\overrightarrow{\sigma} = \sigma_1\mathbf{i}_1 + \sigma_2\mathbf{i}_2 + \sigma_3\mathbf{i}_3,$$

where

$$\sigma_1 = \frac{\partial_1\sqrt{\alpha}}{\sqrt{\alpha}}, \quad \sigma_2 = \frac{\partial_2\sqrt{\beta}}{\sqrt{\beta}}, \quad \sigma_3 = \frac{\partial_3\sqrt{\gamma}}{\sqrt{\gamma}}. \tag{50.19}$$

Theorem 50.12 [11] *Let $\overrightarrow{\mathcal{E}}_1, \overrightarrow{\mathcal{E}}_2$ and $\overrightarrow{\mathcal{E}}_3$ be a set of linearly independent solutions of* (50.18). *Then the quaternionic-valued function*

$$\overrightarrow{\mathcal{E}} = \sum_{k=1}^{3} \varphi_k \overrightarrow{\mathcal{E}}_k$$

will be the general solution of (50.18), *where the real-valued functions φ_1, φ_2 and φ_3 are all solutions of the equation*

$$\sum_{k=1}^{3} (\mathbf{D}\varphi_k)\overrightarrow{\mathcal{E}}_k = 0. \tag{50.20}$$

Indeed, the quaternionic equation (50.18) can be considered a generalization of the Vekua equation (see e.g. [5]), as well as (50.20) could represent a quaternionic generalization of the Beltrami equation (a very interesting work developed from a purely mathematical point of view was posed in [7]).

In order to introduce new solutions for (50.18), we will construct a set of three linearly independent solutions $\overrightarrow{\mathcal{E}}_1, \overrightarrow{\mathcal{E}}_2$ and $\overrightarrow{\mathcal{E}}_3$ for (50.18). Let us assume $\overrightarrow{\mathcal{E}}_1 = \mathcal{E}_1 \mathbf{i}_1$, where $\mathcal{E}_1$ is a real-valued function. Substituting into (50.18) we obtain the differential system

$$\partial_1 \mathcal{E}_1 + \mathcal{E}_1 \sigma_1 = 0, \quad \partial_2 \mathcal{E}_1 - \mathcal{E}_1 \sigma_2 = 0, \quad \partial_3 \mathcal{E}_1 - \mathcal{E}_1 \sigma_3 = 0;$$

for which

$$\mathcal{E}_1 = e^{-\int \sigma_1 dx_1 + \int \sigma_2 dx_2 + \int \sigma_3 dx_3},$$

is a solution. Applying equivalent steps, we can check that

$$\begin{aligned}
\overrightarrow{\mathcal{E}}_1 &= \mathbf{i}_1 \mathcal{E}_1 = \mathbf{i}_1 e^{-\int \sigma_1 dx_1 + \int \sigma_2 dx_2 + \int \sigma_3 dx_3}, \\
\overrightarrow{\mathcal{E}}_2 &= \mathbf{i}_2 \mathcal{E}_2 = \mathbf{i}_2 e^{\int \sigma_1 dx_1 - \int \sigma_2 dx_2 + \int \sigma_3 dx_3}, \\
\overrightarrow{\mathcal{E}}_3 &= \mathbf{i}_3 \mathcal{E}_3 = \mathbf{i}_3 e^{\int \sigma_1 dx_1 + \int \sigma_2 dx_2 - \int \sigma_3 dx_3},
\end{aligned}$$

conform a set of linearly independent solutions of (50.18).

Nonetheless it is not clear how to solve the general case of (50.20) in order to obtain the general solution of (50.18), following [13], we can introduce an infinite set of its solutions bias the following method.

Consider, for example, the particular case of (50.20) when $\varphi_3 = 0$. Thus we have

$$(\mathbf{D}\varphi_1)\overrightarrow{\mathcal{E}}_1 + (\mathbf{D}\varphi_2)\overrightarrow{\mathcal{E}}_2 = 0.$$

Expanding this quaternionic equation, and introducing the notation

$$p = e^{-2\int \sigma_1 dx_1 + 2\int \sigma_2 dx_2},$$

the system takes the form

$$\partial_1 \varphi_1 = -\frac{1}{p^2}\partial_2 \varphi_2, \quad \partial_2 \varphi_1 = \frac{1}{p^2}\partial_1 \varphi_2. \tag{50.21}$$

This differential equations are not other than the so-called p-analytic system [12].

Theorem 50.13 [8] *The real-valued functions φ_1 and φ_2 will be solutions of the system (50.21) if and only if the function*

$$W = \varphi_1 p + i\frac{\varphi_2}{p}$$

is a solution of the Vekua equation

$$\partial_{\overline{\zeta}} W - \frac{\partial_{\overline{\zeta}} p}{p} \overline{W} = 0, \tag{50.22}$$

where $\partial_{\overline{\zeta}} = \partial_2 + i\partial_1$.

This Vekua equation has precisely the form of the equation studied in the previous work, where it was posed the structure of the general solution for the two-dimensional Electrical Impedance Equation.

Noticing that p is in fact a separable-variables function, hence an explicit generating sequence can be introduced in order to express the general solution of (50.22) in terms of Taylor series in formal powers (see e.g. [8], [13])

$$W = \sum_{n=0}^{\infty} Z^n(a_n, \zeta_0; \zeta),$$

where $\zeta = x_2 + ix_1$.

We can use identical procedures considering $\varphi_2 = 0$ in (50.20), and subsequently $\varphi_3 = 0$. This will provide a wider class of new solutions for (50.18) and in consequence for (50.1)

50.3.1 Electrical Current Density Distributions: An Example of σ Depending on x_1

It is evident that every formal power $Z^{(n)}(1, \zeta_0; \zeta)$ and $Z^{(n)}(i, \zeta_0; \zeta)$ will be directly related with a current distribution $\overrightarrow{j}$. Thus, we will consider the case when the electrical conductivity has the form

$$\sigma = e^{2cx_1},$$

where $c \in \mathbb{R}$. For this example, the reader can check that a generating pair corresponding to the Vekua equation (50.22) is

$$F_0 = e^{cx_1}, \quad G_0 = ie^{-cx_1}.$$

The following theorem was originally posed by Bers [3], and latter generalized by Kravchenko [8]. Here we pose it slightly modified for our example [13].

Theorem 50.14 *The generating pair* $F_0 = e^{cx_1}$, $G_0 = ie^{-cx_1}$ *is embedded into a periodic generating sequence with period 1.*

On the light of the last statements, let us consider for example the four first formal powers for the Vekua equation (50.22), considering $\zeta_0 = 0$ [8]:

$$Z^{(0)}(1,0;\zeta) = e^{cx_1}, \quad Z^{(0)}(i,0;\zeta) = ie^{-cx_1};$$

$$Z^{(1)}(1,0;\zeta) = x_2 e^{cx_1} + \frac{i}{c}\sinh(cx_1), \quad Z^{(1)}(i,0;\zeta) = ix_2 e^{-cx_1} - \frac{1}{c}\sinh(cx_1);$$

$$Z^{(2)}(1,0;\zeta) = \left(x_2^2 - \frac{x_1}{c}\right)e^{cx_1} + \frac{1}{c^2}\sinh(cx_1) + i\frac{2x_2}{c}\sinh(cx_1),$$

$$Z^{(2)}(i,0;\zeta) = -\frac{2x_2}{c}\sinh(cx_1) + i\left(\left(x_2^2 + \frac{x_1}{c}\right)e^{-cx_1} - \frac{1}{c^2}\sinh(cx_1)\right).$$

They are all (F_0, G_0)-pseudoanalytic, thus their corresponding electrical current densities can be constructed as follows:

$$\overrightarrow{j}^{(n)} = \begin{pmatrix} \sqrt{\sigma}\mathcal{E}_1 p^{-1}\,\mathrm{Re}\,(Z^{(n)}(1,0;\zeta) + Z^{(n)}(i,0;\zeta)) \\ \sqrt{\sigma}\mathcal{E}_2 p\mathrm{Im}\,(Z^{(n)}(1,0;\zeta) + Z^{(n)}(i,0;\zeta)) \\ 0 \end{pmatrix}.$$

More details about the electrical current patches obtained bias this method, as well as some electric potential in exact form, can be found in [13].

50.4 Conclusions

The study of the Electrical Impedance Equation 50.1 is the base for well understanding the Electrical Impedance Tomography Problem. In this context, the possibility of approaching the general solution of (50.1) by means of the formal powers, opens a new patch for exploring the behavior of the electric potential u at the boundary Γ, because such potential which can be approached by standard numerical methods once the formal powers are obtained (see [6]).

This implies that we can study the behavior of the potential u corresponding to every formal power individually, and then the possibility of suggesting patterns will considerably increase when observing variations of σ inside the domain Ω. Perhaps this point of view will allow us to consider the Electrical Impedance Tomography problem more stable, and in consequence to approach in a more efficient way its solutions.

Acknowledgments The authors would like to acknowledge the support of CONACyT project 106722, Mexico

References

1. Astala K, Päivärinta L (2006) Calderon's inverse conductivity problem in the plane. Ann Math 163:265–299
2. Bayford R, Kantartzis P, Tizzard A, Yerworth R, Liatsis P, Demosthenous A (2007) Reconstruction algorithms to monitor neonate lung function. In: 13th international conference

on electrical bioimpedance and the 8th conference on electrical impedance tomography 2007, IFMBE proceedings, vol 17. Springer, Berlin
3. Bers L (1953) Theory of pseudoanalytic functions. IMM, New York University, New York
4. Calderon AP (1980) On an inverse boundary value problem. Seminar on numerical analysis and its applications to continuum physics, Soc Brasil Mat, pp 65–73
5. Castañeda A, Kravchenko VV (2005) New applications of pseudoanalytic function theory to the Dirac equation. J Phys A Math Gen 38:9207–9219
6. Castillo R, Kravchenko VV, Resendiz R (2010) Solution of boundary value and eigenvalue problems for second order elliptic operators in the plane using pseudoanalytic formal powers. Analysis of PDEs (math.AP); Mathematical Physics (math-ph); Complex Variables (math.CV); Numerical Analysis (math.NA), Available in electronic format in arXiv:1002.1110v1 [math.AP]
7. Kähler U (2000) On quaternionic Beltrami equations. Clifford algebras and their applications in mathematical physics, vol 2. Clifford analysis. Birkhäuser, Basel
8. Kravchenko VV (2009) Applied pseudoanalytic function theory. Series: Frontiers in mathematics, ISBN: 978-3-0346-0003-3
9. Kravchenko VV (2003) Applied quaternionic analysis researches and exposition in mathematics, vol 28. Heldermann Verlag, Lemgo
10. Kravchenko VV, Oviedo H (2007) On explicitly solvable Vekua equations and explicit solution of the stationary Schrödinger equation and of the equation $\mathrm{div}(\sigma\nabla u) = 0$. Complex Variables Elliptic Equations 52(5):353–366
11. Kravchenko VV, Ramirez MP (2010) On Bers generating functions for first order systems of mathematical physics. Advances in Applied Clifford Algebras, ISSN 0188-7009, accepted for publication
12. Polozhy GN (1965) Generalization of the theory of analytic functions of complex variables: p-analytic and (p,q)-analytic functions and some applications. Kiev University Publishers, Kiev (in Russian)
13. Ramirez MP (2010) On the electrical current distributions for the generalized Ohm's Law. Available in electronic form in arxiv.org, 2010
14. Ramirez MP, Gutierrez JJ, Sanchez VD, Bernal F (2010) New solutions for the three-dimensional electrical impedance equation and its application to the electrical impedance tomography theory. In: Proceedings of the World Congress on Engineering 2010, vol I, pp 527–532
15. Ramirez MP, Sanchez VD, Rodriguez O, Gutierrez A (2010) On the general solution for the two-dimensional electrical impedance equation in terms of Taylor series in formal powers. IAENG Int J Appl Math 39(4), IJAM_39_4_13, vol 39, issue 4, ISSN: 1992-9986 (online version); 1992-9978 (print version), accepted for publication
16. Vekua IN (1962) Generalized analytic functions international series of monographs on pure and applied mathematics. Pergamon Press, London

Chapter 51
Modelling of Diseased Tissue Diffuse Reflectance and Extraction of Optical Properties

Shanthi Prince and S. Malarvizhi

Abstract The measurement of optical properties of biological tissues remains a central problem in the field of biomedical optics. Knowledge of these parameters is important in both therapeutic and diagnostic applications of light in medicine. In this work, based on the diffuse reflectance spectra obtained through experimentation the optical properties are determined. Reflectance spectra are obtained by setting up a simple diffuse reflectance spectroscopic system based on fiber optics. Light from a white light source is incident onto the tissue site through the reflectance probe and the backscattered light is collected by the same and impinged on a spectrometer which generates a spectrograph which is acquired in the system. Empirical forward light transport model based on diffusion theory is formulated. The reflectance spectra obtained is fitted to the proposed spectral (analytical) model. The model coefficients determined by non-linear least square optimization method are used to determine the optical properties of the tissue.

51.1 Introduction

Real time analysis of tissue properties namely, absorption coefficient μ_a and reduced scattering coefficient μ_s' by noninvasive methods finds wide variety of applications, particularly in disease diagnosis and to measure tissue metabolic

S. Prince (✉) · S. Malarvizhi
Department of Electronics and Communication Engineering,
SRM University, SRM Nagar, Kattankulathur 603203, Tamil Nadu, India
e-mail: shanthiprince@ktr.srmuniv.ac.in

S. Malarvizhi
e-mail: hod.ece@ktr.srmuniv.ac.in

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*,
Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_51,
© Springer Science+Business Media B.V. 2011

states. Also, it is used to estimate the depth of penetration of light radiation to determine the dosimetry in photodynamic therapy. Methods to accurately determine optical properties can lead to optical diagnostics tools [1], improvements in laser surgery [2], quantitative determination of chromophore [3] and fluorophore [4] concentrations, drug pharmacokinetics and improvements on Photodynamic Therapy (PDT) dosimetry [5].

Experimental determination of tissue optical properties has been proposed using different methodologies. Integrating sphere [6], frequency domain diffuse reflectance [7], time domain diffuse reflectance [8] and spatially resolved steady-state diffuse reflectance [9] are among the most widely used. Each technique has its own advantages and disadvantages. In this work, based on the steady-state diffuse reflectance method the diffuse reflectance spectra are used for the determination of the optical properties. Advantages of using this method are the inexpensive equipment involved and the simplicity of the measurements.

51.2 Instrumentation

In diffuse reflectance spectroscopy, light is delivered to tissue and after several successive scattering and absorption events it re-emerges from the tissue bearing information about tissue underneath. The collection of re-emerged data (diffuse reflectance) and analysis of it, correlating it with the morphology and the biochemical composition play a vital role in diagnostic applications. Hence the practical means of analyzing the interaction of light with the biological tissues is to monitor the light that has diffusely reflected from the surface. The reflectance further varies in subjects of different complexion and body structure.

Light reflected from a surface consists of specularly reflected and diffusely reflected components. The intensity of the specular component is largely determined by the surface properties of the sample. The intensity of the diffuse component, which includes the contributions from the absorbance and the scattering of light by the specimen and the substrate, can be used to determine the concentration of the indicator species [10].

The spectral reflectance of the tissue conveys information about the metabolites which constitute the tissue. The change in the amount of chromophores in the tissue can be predicted by means of modified Beer Lambert Law [11]. The quantification of chromophores has the potential to provide beneficial information for diagnostic and therapeutic decision making.

Basically, a spectrophotometric measurement is to be done so that, the obtained spectra can be related to the molecular composition and structure of biochemical species in the sample of interest. The schematic diagram of the visible/near-infrared (VIS–NIR) spectroscopy system is shown in Fig. 51.1. It consists of tungsten halogen light source (LS-1, Ocean Optics, Inc., www.oceanoptics.org) which is a versatile white-light source optimized for the VIS–NIR (360–2500 nm) wavelength range, fiber optic reflectance probe (R400, Ocean Optics, Inc., www.oceanoptics.org),

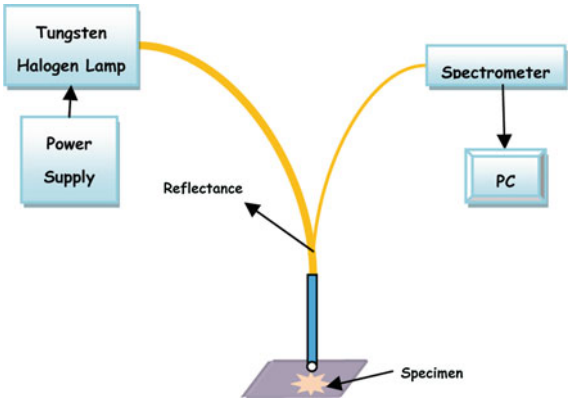


Fig. 51.1 Schematic diagram of diffuse reflectance spectroscopic system

spectrometer (USB4000, Ocean Optics, Inc., www.oceanoptics.org) with CCD device, and a computer with an acquisition software (Spectra Suite, Ocean Optics, Inc., www.oceanoptics.org). The USB4000 is responsive from 200 to 1100 nm, with an optical resolution of ~0.3 nm (FWHM). It consists of 3648 element detector with shutter, high speed electronics and interface capabilities. The reflectance probe (R400) consists of bundle of seven optical fibers. Six illumination fibers and one read fiber—each of which is 400 μm in diameter. The fiber ends are coupled in such a manner that the six-fiber leg (the illumination leg) is connected to the light source and the single fiber leg is connected to the spectrometer. The details of this diffuse reflectance spectroscopic system and data acquisition are elaborated elsewhere [12, 13].

Also, the instrument is subjected to gauge repeatability and reproducibility test, the results of which states that its reproducibility and repeatability is well within the acceptable range.

51.3 Acquisition of Spectra

Initially the reference and dark spectra measurements to correct for instrument response variables are obtained and stored. Reference spectrum is obtained from $BaSO_4$ diffuse reflectance standard. Dark spectrum is obtained with no input light to the spectrometer. To acquire data from the sample tissue, white light from a tungsten halogen lamp is brought to the skin via the reflectance probe. The light penetrates through the skin. Water, hemoglobin species, cytochromes, lipids and proteins absorb this light at specific frequencies. The remaining light is scattered by the skin, with some light being scattered back to the fiber optic probe. The collector fiber in the reflectance probe collects the diffused light from the skin, and directs it to the spectrometer. The spectrometer measures the amount of light and transforms the data collected into digital information. This acquired sample information in the digital form is passed to SpectraSuite (Data acquisition

software). Each recorded spectrum represented the average of forty spectra with 100 ms integration time.

In this experimental setup, light is delivered using fiber reflectance probe with center-to-center separation of about 400 μm at the distal end of the probe, which is placed above the tissue. Prior to data acquisition from normal and diseased skin, a reference spectrum is acquired on a $BaSO_4$ diffuse reflectance standard and all skin spectra are subsequently divided by this reference spectrum for normalization. This normalization accounts for day-to-day variations in the wavelength and magnitude dependence of the light source and detector sensitivity. Before obtaining the readings, the subject's skin and the end of the probe are cleansed with 70% alcohol.

For the present study, zero melanin condition vitiligo and thrombus due to injury are chosen. Patients with the above mentioned skin disorders are chosen and their consent to take part in this study is obtained. Selection of the patients of either sex is basically restricted to the age group of 30–40 years of age. From each patient, control spectrum (from normal skin) and condition spectrum (from disease afflicted skin) are obtained. In our previous work [14] are shown the in vivo diffuse reflectance intensity spectra for vitiligo and thrombus along with the normal (control) spectra for each case using the experimental set up described here.

51.4 Empirical Forward Light Transport Model

To extract the useful information from the experimentally acquired reflectance spectra there are two approaches—analytical approach and diagnostic approach. The analytical approach seeks to deconstruct the reflectance spectra into biologically and physically meaningful factors. The diagnostic approach is based upon statistical metrics that separate the normal and disease tissue states.

Re-emerged reflectance spectra bear information about the tissue. In analytical approach a spectral model is proposed based on the light transport through tissue which relies on the spectral characteristics of the tissue chromophores for the determination of the absorption coefficient and on a simple wavelength dependent expression for the determination of the reduced scattering coefficient.

The proposed spectral model is fitted into the experimentally obtained normalized reflectance spectra and model coefficients are estimated. Using the model coefficients in the proposed spectral model, optical properties are calculated.

Reflectance measurements on tissue samples $R_s(\lambda)$ are normalized by the reflectance from barium sulphate ($BaSO_4$) standard $R_{\text{std}}(\lambda)$ and the final spectrum is denoted as the ratio R_{norm}:

$$R_{\text{norm}}(\lambda) = \frac{R_{\text{s}}(\lambda)}{R_{\text{std}}(\lambda)} = \frac{S(\lambda)T_{\text{s}}(\lambda)\eta_{\text{c,s}}(\lambda)D(\lambda)}{S(\lambda)T_{\text{std}}(\lambda)\eta_{\text{c,std}}(\lambda)D(\lambda)} \tag{51.1}$$

where $S(\lambda)$ is the light source power, $D(\lambda)$ is the detector sensitivity, $T_s(\lambda)$ is the optical transport into the sample medium and returning to the sample surface at the collection fiber, $T_{\text{std}}(\lambda)$ is the optical transport into the standard medium and

returning to the surface at the collection fiber, $\eta_{c,s}(\lambda)$ is the optical fiber collection efficiency for the sample and $\eta_{c,std}(\lambda)$ is the optical fiber collection efficiency for the standard.

The terms S (the source spectral response) and D (the detector spectral response) are the same for samples and standard measurements and does not vary within a measurement procedure and thus cancel in Eq. 51.1. The normalized measurement, $R_{norm}(\lambda)$ can be written as

$$R_{norm}(\lambda) = \frac{T_s(\lambda)\eta_{c,s}(\lambda)}{T_{std}(\lambda)\eta_{c,std}(\lambda)} \tag{51.2}$$

$$R_{norm}(\lambda) = K * T_s(\lambda) \tag{51.3}$$

the factor K is a constant given by $K = \frac{\eta_{c,s}(\lambda)}{T_{std}(\lambda)\eta_{c,std}(\lambda)}$.

To model the tissue reflectance the scattering behaviour and the absorption behavior in the tissue should be modeled. The scattering behaviour is dominated by Rayleigh scattering $\mu_s'(\lambda_{ray})$ at short wavelengths below 650 nm from collagen fibril fine structure, small membranes, and other ultra structure on the 10–100 nm scale [15]. The Mie scattering $\mu_s'(\lambda_{mie})$ is dominating for wavelengths above 650 nm from larger tissues structures such as collagen fiber bundles, mitochondria, nuclei and cells. Therefore, the visible to near-infrared spectral region is significantly affected by both scattering that is given by

$$\mu_s'(\lambda) = a\left(\mu_s'(\lambda_{mie}) + \mu_s'(\lambda_{ray})\right) \tag{51.4}$$

where "a" [dimensionless] is the factor that characterizes magnitude of scattering, $\mu_s'(\lambda_{mie})$ and $\mu_s'(\lambda_{ray})$ [16]. The scattering properties are specified by fitting for scattering factor "a".

Tissue absorption is modelled as a linear combination of absorption coefficient of water (μ_{water}) with a volume fraction of f_w, a background spectrum for skin i.e. absorption coefficient of melaninless hemoglobin-water free tissue (μ_{a_other}), volume fraction of melanosomes (f_{mel}) and a variable blood volume fraction (f_b) of oxygenated and deoxygenated whole blood (μ_{a_oxy}—the absorption coefficient of oxy-hemoglobin, μ_{a_deoxy}—the absorption coefficient of deoxy-hemoglobin) at an oxygen saturation (S). The water content is fixed as 75%. In principle, the water content could be fitted, but the system is not sufficiently sensitive in the 900–1000 nm spectral region, where water strongly influences the spectra. In the framework of this model, the absorption coefficients of dermal layers taking into account the spatial distribution of melanosomes, blood and water content within the skin is written as:

$$\begin{aligned}\mu_a(\lambda) = {} & f_b\left[S \cdot \mu_{a_{oxy}} + (1-S)\mu_{a_{deoxy}}\right] + f_w \cdot \mu_{water} + (1 - f_{mel}) * (1 - f_b) \\ & * (1 - f_w) * \mu_{a\_other}\end{aligned} \tag{51.5}$$

where $\mu_{a\_other}[cm^{-1}]$ the absorption coefficient of melaninless hemoglobin-water free tissue given by $\mu_{a\_other} = 7.84 \times 10^8 \times \lambda^{-3.255}\,cm^{-1}$ [16] and μ_{water} $[cm^{-1}]$ is

the absorption coefficient of water. The absorption coefficient of water is obtained from the extinction spectra of water by Hale and Querry [17]. For oxyhemoglobin and deoxyhemoglobin it is obtained from their extinction spectra according to the Takatani and Graham [18].

In this work spatially resolved steady-state diffuse reflectance model, where the source detector separation of 0.4 mm is used for the determination of optical properties. To analyze the reflectance measurements, appropriate light transport model is to be employed. The light transport model helps determine how light from the source fiber reaches the collection fiber. For this, a number of analytical expressions have been proposed [19]. In this work the Farrell model for pencil beam irradiance [9] is chosen. In this model, the deeper dermis layer is assumed to be infinitely thick, implicitly assuming subcutaneous tissue, such as fat and muscle to be of negligible influence on the reflectance spectrum. This is a fair assumption for most visible wavelengths. The fraction of transport T collected by fiber at a radial distance "r" from the source is given by

$$T\left(\mu_{\mathrm{a}}, \mu_{\mathrm{s}}'\right) = \frac{1}{4\pi}\left[Z_{\mathrm{o}}\left[\mu_{\mathrm{eff}} + \frac{1}{r_1}\right]\frac{\mathrm{e}^{-\mu_{\mathrm{eff}}\cdot r_1}}{r_1^2} + \left(Z_0 + 2Z_{\mathrm{b}}\right)\left[\mu_{\mathrm{eff}} + \frac{1}{r_2}\right]\frac{\mathrm{e}^{-\mu_{\mathrm{eff}}\cdot r_2}}{r_2^2}\right] \quad (51.6)$$

where $Z_{\mathrm{o}} = 1/\left(\mu_{\mathrm{a}} + \mu_s'\right)$, $Z_{\mathrm{b}} = 2AD$, $D = Z_{\mathrm{o}}/3$ is the diffusion constant, $\mu_{\mathrm{eff}} = \left[\sqrt{(D/\mu_{\mathrm{a}})}\right]^{-1}$ is the effective transport coefficient, $r_1 = \sqrt{\left(Z_{\mathrm{o}}^2 + r^2\right)}$, $r_2 = \sqrt{\left(Z_{\mathrm{o}} + 2Z_{\mathrm{b}}\right)^2 + r^2}$ and $A = (1 + r_i)/(1 - r_i)$.

The term r_i is the internal specular reflection parameter due to the refractive index mismatch at the surface given as

$$r_i = 0.668 + 0.0636n + \frac{0.710}{n} - \frac{1.440}{n^2},$$

n is the refractive index of the tissue.

Spectral model for light transport based on diffusion theory is formulated. The predicted reflectance is given as

$$R_{\mathrm{p}} = G * T\left(\mu_{\mathrm{a}}, \mu_{\mathrm{s}}'\right) \quad (51.7)$$

where T is the fraction of transport collected by fiber at a radial distance "r" from the source and G is the proportionality factor which contains factors like the fiber diameter, numerical aperture, and the ratio between the optical fiber probe collection efficiency for the sample and the standard.

For the case of skin studies the light transport in epidermis is modeled as

$$T_{\mathrm{epi}} = \mathrm{e}^{-f_{\mathrm{mel}}\mu_{\mathrm{a\_mel}}L_{\mathrm{epi}}} \quad (51.8)$$

where $\mu_{\mathrm{a\_mel}}$ [cm^{-1}] is the absorption coefficient of melanin defined as $\mu_{\mathrm{a_{mel}}}$ [17], L_{epi}, equals twice the epidermal thickness. The volume fraction of melanin f_{mel} in typical 60 μm thick epidermis is specified by fitting Eq. 51.8.

The model coefficients a, S, f_b and f_{mel} are estimated by fitting the predicted reflectance measurements to the obtained experimental measurements by non-linear least square optimization method. Substituting the model coefficient values in Eqs. 51.4 and 51.5 optical properties of the diseased tissue are determined.

51.5 Results and Discussion

Once the model coefficients are estimated, then using Eqs. 51.4 and 51.5, we calculate the absorption coefficient $\mu_a(\lambda)$ and the reduced scattering coefficient $\mu_s'(\lambda)$ of the different diseased tissue under study. The knowledge of the scattering properties can reveal information about the morphology and architecture of skin [20]. The knowledge of absorption properties can reveal information about the biochemistry.

For the validation of the model, measurements were performed on the human forearm and their optical properties estimated. Figure 51.2 shows the result of the spectral model study on forearm with plots of the experimental and predicted spectra. The percentage residual error [(predicted − experimental)/experimental] is shown in the bottom plot. The plot of the obtained reduced scattering coefficients and the absorption coefficients for the forearm as a function of wavelength are shown in Fig. 51.3.

The result of this work is compared with other similar works. Table 51.1 shows the comparison of $\mu_a(\lambda)$ and $\mu_s'(\lambda)$ values for forearm obtained by this method with

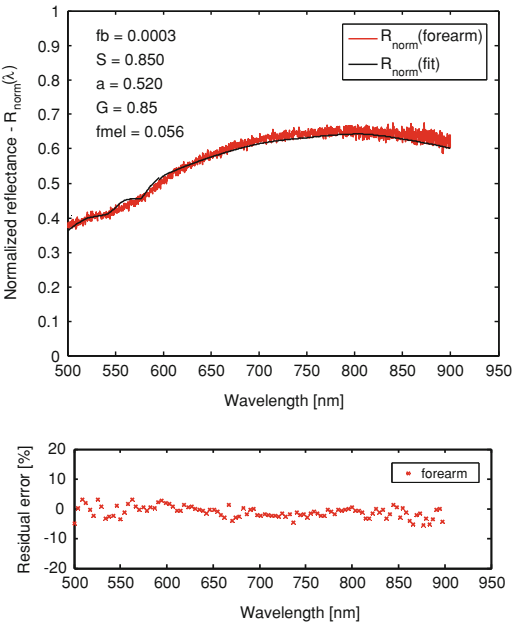


Fig. 51.2 Normalized reflectance spectra of forearm in comparison with the predicted values determined using the fitted parameters a, S, G, f_{mel} and f_b. *Bottom plot* shows the percentage residual errors [(predicted − measured)/measured × 100%]

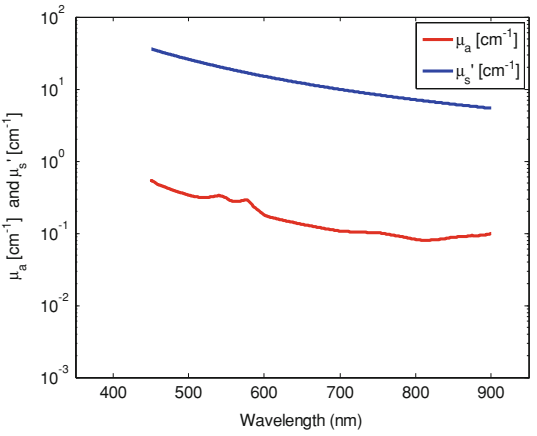


Fig. 51.3 Absorption coefficients (μ_a) and reduced scattering coefficients (μ_s') of forearm determined from the spectral model

Table 51.1 Comparison of $\mu_a(\lambda)$ and $\mu_s'(\lambda)$ for forearm obtained by the proposed spectral model with the results from literature

Wavelength λ (nm)	Results from this model		Results from other related works		Literature references
	μ_a (cm^{-1})	μ_s' (cm^{-1})	μ_a (cm^{-1})	μ_s' (cm^{-1})	
633	0.14	13.0	0.17	9.08	[21]
660	0.1269	11.61	0.128	8.68	[21]
700	0.107	9.948	0.09	8.1	[21]
750	0.1018	8.349	0.03	10	[22, 23]

that of other works in literature, with the corresponding references. It is clear from Table 51.1 that the results obtained through this model agree well with those of Doornbos [21] as compared with others. The discrepancy in $\mu_s'(\lambda)$ values may be because of the difference of individuals, including skin color and age.

Figure 51.4a, b show the normalized reflectance spectra for vitiligo and thrombus in comparison to the predicted values determined using the fitted parameters a, S, G, f_{mel} and f_b. The model coefficient values of a, S, G, f_{mel} and f_b after best fit are specified in the figures. Bottom curves show the percentage residual errors [(predicted − measured)/measured × 100%].

Table 51.2 shows the values of fitted parameters a, S, G, f_{mel} and f_b, for normal (control skins) sites and diseased sites.

On analyzing the values it is observed that volume fraction of blood is a little higher for thrombus as there is a large amount of blood accumulation due to clot. Observing the oxygen saturation, the average value is around 85% as against the standard value of 90–98%. It may be due to the fact that the estimation includes both arterial and venous blood. As expected the volume fraction of melanosomes is higher for thrombus, since the level of pigmentation (due to blood clot) is higher compared to the other sites. However, for the vitiligo site as it is the region of zero melanin the value is the lowest.

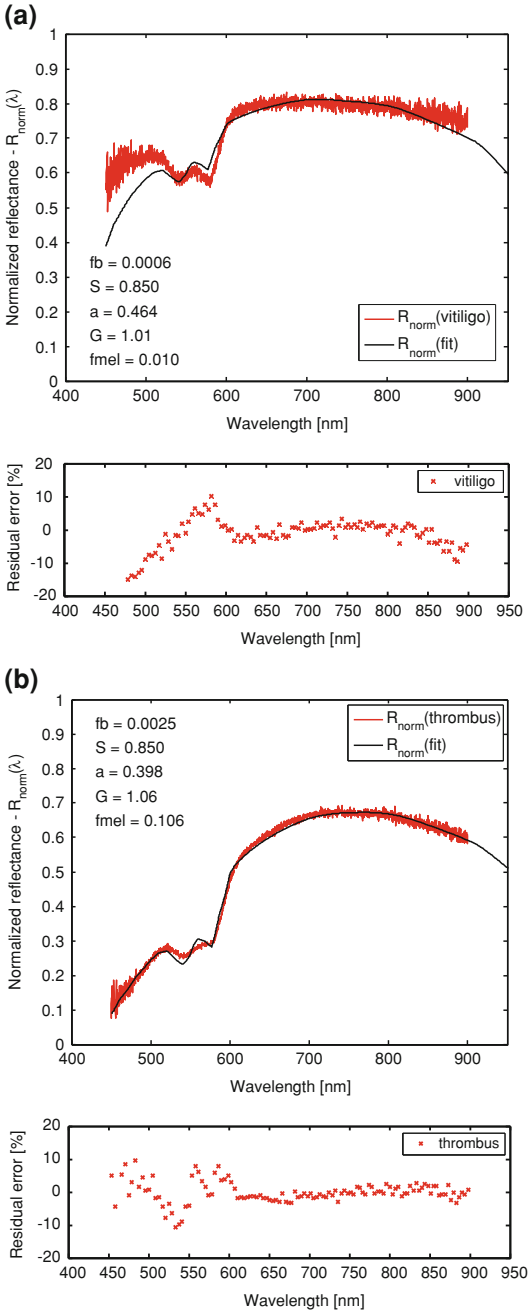


Fig. 51.4 Normalized reflectance spectra for **a** vitiligo and **b** thrombus in comparison with the predicted values determined using the fitted parameters a, S, G, f_{mel} and f_b. *Bottom plot* shows the percentage residual errors [(predicted − measured)/measured × 100%]

Table 51.2 Values of fitted parameters a, S, G, f_{mel} and f_b for normal (control skin) sites and diseased sites

Skin type	a (–)	S (%)	G (–)	f_{mel} (%)	f_b^a (%)
Control skin #1	0.451	85	0.72	1.5	0.9
Control skin #2	0.405	96	1.46	13	0.1
Vitiligo	0.464	85	1.01	1	0.6
Thrombus	0.398	85	1.06	10	2.5

[a] The values displayed in the figures are scaled

From the figures it is clear that the results of this model agree well with the experimental results. Overall the predicted data fits well above 600 nm as seen from the residual error displayed at the bottom of each graph. The residual error is less than 5% for spectral region from 650 to 850 nm. Disagreement is noticeable in the fitting of spectra below 600 nm, as expected since diffusion theory fails when the reduced mean free path $(1/(\mu_a + \mu_s'))$ is comparable with the source-detector separation and when μ_a is comparable to μ_s'.

Figure 51.5a, b shows the plot of absorption and reduced scattering coefficients of vitiligo skin and thrombus in comparison with their control skin. From the vitiligo optical properties plot it can be observed that the absorption coefficient shows changes between normal and vitiligo site in the visible wavelength range of 400–600 nm, with higher values for the vitiligo skin. The reduced scattering coefficient values are higher for the vitiligo throughout the spectral region under study.

The plot of absorption and reduced scattering coefficients for the thrombus tissue in comparison with the normal tissue is shown in Fig. 51.5b. Here the reduced scattering coefficient and absorption coefficient for the thrombus tissue is higher when compared to the normal tissue. Thrombus being the region of blood clot, the volume fraction of hemoglobin is higher leading to the contribution towards the higher values for optical coefficients.

51.6 Conclusion

We have presented an analytical model for the study of diffuse reflectance spectra from normal and diseased tissue. The model provides quantitative information for the hemoglobin and melanin contents, which are the dominant chromophores of skin in the visible range. In addition, it provides information about the light scattering and absorption properties of the skin. The technique described here is good for the analysis and evaluation of human skin in vivo, in a rapid and non-invasive way, and the information obtained is potentially useful in the assessment and diagnosis of numerous pathologic conditions. Finally, this diffuse reflectance model is based on certain assumptions that the absorption is due to only melanin, water and hemoglobin. There are other chromophores that have specific absorption spectrum in the skin that dominates during certain pathological conditions.

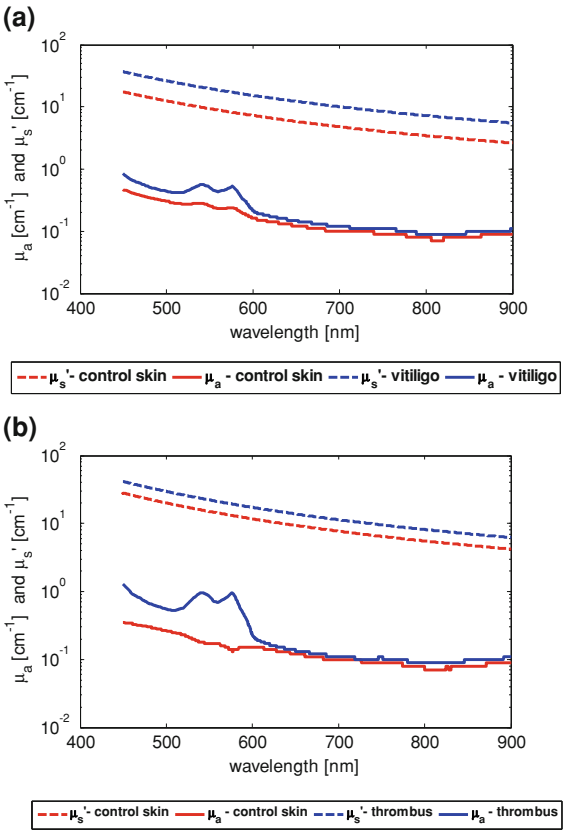


Fig. 51.5 Absorption coefficients (μ_a) and reduced scattering coefficients (μ_s') determined for **a** vitiligo from the spectral model, **b** thrombus by the spectral model

In order to cover wide pathological study this model can be remodeled to incorporate other chromophores.

Acknowledgments This work is being funded and supported by All India Council for Technical Education (AICTE), Government of India, New Delhi under the scheme of Career Award for Young Teacher (CAYT).

References

1. Peters VG, Wyman DR, Patterson MS, Frank L (1990) Optical properties of normal and diseased human breast tissues in the visible and near infrared. Phys Med Biol 35:1317–1334
2. Jacques SL (1992) Laser-tissue interactions: photochemical, photothermal, photomechanical. Surg Clin N Am 72:531–558
3. Patterson MS, Schwartz E, Wilson BC (1989) Quantitative reflectance spectrophotometry for the noninvasive measurement of photosensitizer concentration in tissue during photodynamic therapy. In: Dougherty TJ (ed) Photodynamic therapy: mechanisms, Proc SPIE, vol 1065, pp 115–122

4. Gardner CM, Jacques SL, Welch AJ (1996) Fluorescence spectroscopy of tissue: recovery of intrinsic fluorescence from measured fluorescence. Appl Opt 35:1780–1792
5. Jacques SL (1989) Simple theory, measurements, and rules of thumb for dosimetry during photodynamic therapy. In: Photodynamic therapy: mechanisms, T. J. Dougherty, Proc SPIE, vol 1065, pp 100–108
6. Pickering JW, Moes CJM, Sterenborg HJCM, Prahl SA, van Gemert MJC (1992) Two integrating spheres with an intervening scattering sample. J Opt Soc Am A 9:621–631
7. Fantini S, Francechini-Fantini MA, Maier JS, Walker SA, Barbieri B, Gratton E (1995) Frequency-domain multichannel optical detector for noninvasive tissue spectroscopy and oximetry. Opt Eng 34:32–42
8. Patterson MS, Chance B, Wilson BC (1989) Time resolved reflectance and transmittance for the non-invasive measurements of optical properties. Appl Opt 28:2331–2336
9. Farrel TJ, Patterson MS (1992) A diffusion theory model of spatially resolved, steady-state diffuse reflectance for the noninvasive determination of tissue optical properties in vivo. Med Phys 19:879–888
10. Vo-Dinh T (2003) Biomedical photonics handbook. CRC Press, Boca Raton
11. Tuchin V (2007) Tissue optics light scattering methods and instruments for medical diagnosis. SPIE Press, Bellingham
12. Prince S, Malarvizhi S (2008) Functional optical imaging of a tissue based on diffuse reflectance with fibre spectrometer. In: 4th European congress for medical and biomedical engineering, IFMBE proceedings, vol 22, pp 484–487
13. Prince S, Malarvizhi S (2008) Multi wavelength diffuse reflectance plots for mapping various chromophores in human skin for non-invasive diagnosis. In: 13th international conference on biomedical engineering, IFMBE proceedings, vol 23, pp 323–326
14. Prince S, Malarvizhi S (2010) Estimation of optical properties of normal and diseased tissue based on diffuse reflectance spectral model. Lecture notes in engineering and computer science: Proceedings of the world congress on engineering WCE 2010, 30 June–2 July, 2010, London, UK, pp 578–582
15. Jacques SL (1996) Modeling light transport in tissue. In: Verga Scheggi AM et al (eds) Biomedical optical instrumentation and laser-assisted biotechnology. Kluwer Academic Publishers, Dodrecht, pp 21–32
16. Jacques SL, Skin optics. http://omlc.ogi.edu/news/jan98/skinoptics
17. Hale GM, Querry MR (1973) Optical constants of water in the 200 nm to 200 μm wavelength region. Appl Opt 12:555–563
18. Takatani S, Graham MD (1987) Theoretical analysis of diffuse reflectance from a two-layer tissue model. IEEE Trans Biomed Eng BME-26:656–664
19. Kienle A, Patterson MS (1997) Improved solutions of the steady-state and the time-resolved diffusion equations for reflectance from a semi-infinite turbid medium. J Opt Soc Am A 14(1):246–254
20. Ferdman AG, Yannas IV (1993) Scattering of light from histologic section: a new method for the analysis of connective tissue. J Invest Dermatol 100:710–716
21. Doornbos RMP, Lang R, Aalders MC, Cross FW, Sterenborg HJCM (1999) The determination of in vivo human tissue optical properties and absolute chromophore concentrations using spatially resolved steady-state diffuse reflectance spectroscopy. Phys Med Biol 44:967–981
22. Farrell TJ, Patterson MS, Essenpreis M (1998) Influence of layered tissue architecture on estimates of tissue optical properties obtained from spatially resolved diffuse reflectometry. Appl Opt 37:1958–1972
23. Nickell S, Hermann M, Essenpries M, Farrell TJ, Kramer U, Patterson MS (2000) Anisotropy of light propagation in human skin. Phys Med Biol 45:2873–2886

Chapter 52
Vertical Incidence Increases Virulence in Pathogens: A Model Based Study

Priti Kumar Roy, Jayanta Mondal and Samrat Chatterjee

Abstract Two model systems are proposed and analyzed. In the first system the disease transmission in the population is considered to be purely horizontal, while in the second system the transmission is considered to be mixed, i.e., both horizontal and vertical transmissions are present in the system. Analytically both the systems are analyzed separately and then numerical simulations are performed to observe the effect of disease transmission in disease dynamics. Our results showed that the presence of vertical transmission along with the horizontal transmission increases the virulence of the pathogens.

52.1 Introduction

Host–pathogen models are mathematical prototypes pertaining to epidemiology and these are of immense importance in view of the emergence and re-emergence of epidemiological diseases in the present day global scenario. This kind of study

P. K. Roy (✉)
Centre for Mathematical Biology and Ecology, Department of Mathematics, Jadavpur University, Kolkata, 700032, India
e-mail: pritiju@gmail.com

J. Mondal
Department of Mathematics, Barasat College, Kolkata, 700124, India
e-mail: mjayanta06@yahoo.in

S. Chatterjee
Immunology Group, International Centre for Genetic Engineering and Biotechnology, Aruna Asaf Ali Marg, New Delhi, 110067, India
e-mail: samrat_ct@rediffmail.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_52,
© Springer Science+Business Media B.V. 2011

has its own importance in the field of biology because pathogens are a diverse and deadly group of organisms whose members include the causative agents of AIDS, malaria and tuberculosis. These microbial invaders have evolved complex strategies that enable them to proliferate within animal cells and evade many host defences [1]. In terms of population dynamics, the presence of infection divide the host populations into two classes, susceptible or healthy class and infected individuals, while pathogens are assumed to cause infection which transforms susceptible individuals to infected individuals. A large part of current deterministic epidemic theory extend this idea which was first proposed by Ross [2, 3]. After that many researchers had used host–pathogen models to study other aspects of species interactions, for example see [4–10]. Some of these models also include vaccination. Most of these models assumed that infected individuals are formed from susceptible class and that they cannot reproduce. But in reality it has been seen that infected individuals give birth. In some cases they give birth to healthy offsprings, while in other cases the offsprings are also infected. For example, individuals infected with diseases like influenza [11] and athlete's foot Impetigo [12] can give birth to healthy offsprings which are susceptible to that diseases. In other diseases like HIV [13], Hepatitis B [14] and genital wart [15] the infected individual give birth to a individual infected with the same disease. In either case the infected individual has the potential to reproduce and hence it becomes important to consider the reproduction of the infected individual while making host pathogen models.

In the present study we considered the reproduction ability of the individuals and focused our study on the role of disease transmission. There are articles where the authors work on the pathogen transmission in host–pathogen model [16]. They considered different types of transmissions between susceptible and infected hosts, restricting their study to horizontal transmission only. But here we are interested to see how presence of vertical transmission changes the virulence of a disease. The idea that vertical transmission of parasites selects for lower virulence is widely accepted [17]. The prediction that vertical transmission alone should select for decreased virulence, compared with horizontal transmission, has been confirmed experimentally in a bacterial host–symbiont system, using a hybrid bacteriophage-plasmid as a symbiont [18]. Although it is clear that vertical transmission alone selects for decreased virulence of symbiont/ parasites, the case of most interest is that of mixed horizontal and vertical transmission [17]. Ewald [19] has proposed that infections with mixed modes of transmission should fall along a continuum, with mostly horizontally transmitted parasites tending toward higher virulence, and mostly vertically transmitted agents tending toward lower virulence.

In this article we have extended our previous work [4]. In that research article it was consider a conventional host–pathogen model including a recovery of the infected individuals to the healthy organisms termed as susceptible. Our analysis showed that the removal of infected host population is caused by the biological and physical realizable threshold of recovery rate. Here we proposed two systems—in one the transmission is purely horizontal (though the infected population still reproduce) and in the other system the transmission is mixed, i.e., both

horizontal and vertical transmissions are present. We want to see the difference in the dynamical nature of the two systems emphasizing on the disease extinction and persistence criteria. The main question here is—whether the vertical transmission increase virulence or not?

52.2 Infected Class Giving Birth to Susceptible Offspring

A host pathogen model consisting of a host population, whose concentration is denoted by N ($[N]$ = number of host per designated area) and a pathogen population inflicting infection in the host population whose concentration is denoted by V ($[V]$ = number of pathogens per designated area) is considered. In the presence of pathogenic infection, the host population is divided into two disjoint classes, susceptible S, and infected I.

The following assumptions are made to formulate the basic model equations.

(A1) In the absence of pathogen the growth of the host population follows the logistic law implying that this growth is entirely controlled by an intrinsic birth rate constant $r(\in R_+)$ with a carrying capacity $K(\in R_+)$. The mathematical form of such logistic growth is

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right). \tag{52.1}$$

(A2) Introduction of pathogen in the system splits the host population into two disjoint classes, namely susceptible host S and infected host I, such that at any time t the total host population remains as

$$N(t) = S(t) + I(t). \tag{52.2}$$

(A3) Susceptible is assumed to produce susceptible offspring and so S increases its population by reproduction as per logistic law (52.1). Infected host also give birth to susceptible offspring and so the growth of I is added to the susceptible class.

(A4) It is assumed that the spread of disease takes place in two avenues namely, by pathogens as well as by contact of a susceptible host with a infected host following the law of mass action. It should be noted here that some researchers argued in favor of proportional mixing rate of contact between S and I rather than a simple law of mass action. But Greenwood experiment [20] on prototype systems showed that quantitative results remain the same in either cases of the mentioned contact processes.

Following assumptions (A3) and (A4), Eq. 52.1 can be written as

$$\frac{dS}{dt} = r(S+I)\left(1 - \frac{S+I}{K}\right) - \lambda SI - \gamma SV, \tag{52.3}$$

where $\lambda(\in R_+)$ is the intensity of infection by infected host and $\gamma(\in R_+)$ is the force of infection through contact with pathogens. The equation depicting the dynamics of pathogen population thus becomes

$$\frac{dV}{dt} = -\gamma SV + \eta d_I I - \mu V, \tag{52.4}$$

where $d_I(\in R_+)$ is the death rate constant of I. Note that we consider the mortality of I to be completely due to lysis and there exists no separate base line mortality of it. Here $\eta(\in R_+)$ is the rate of cell lysis (replication of pathogens) and the natural death rate of pathogens is denoted as $\mu(\in R_+)$.
Based on the string of arguments the time rate of change of I can be written as

$$\frac{dI}{dt} = \lambda SI + \gamma SV - d_I I. \tag{52.5}$$

(A5) We assumed that the infected hosts can recover from pathogenic infection and move to add to the susceptible host population. Such recovery would stem out from immunization or vaccination. We consider a recovery rate of I to be denoted by $\delta(\in R_+)$.
Following the above assumption (A1)–(A5), the final set of equations depicting the dynamics of susceptible host, infected host and pathogens can be written as

$$\begin{aligned}
\frac{dS}{dt} &= r(S+I)\left(1 - \frac{S+I}{K}\right) - \lambda SI - \gamma SV + \delta I,\\
\frac{dI}{dt} &= \lambda SI + \gamma SV - d_I I - \delta I,\\
\frac{dV}{dt} &= -\gamma SV + \eta d_I I - \mu V,
\end{aligned} \tag{52.6}$$

where, $S(0) > 0,\ I(0) > 0,\ V(0) > 0$.

The variables of model equations (52.6) are made dimensionless by rescaling all the variables in terms of carrying capacity K. Thus we apply the transformation, $s = \frac{S}{K},\ i = \frac{I}{K},\ v = \frac{V}{K},\ \tau = rt$ and get the following dimensionless form of the model equation (52.6). For notational convenience we will replace τ by t

$$\begin{aligned}
\frac{ds}{dt} &= (s+i)(1-(s+i)) - asi - bsv + ci,\\
\frac{di}{dt} &= asi + bsv - di - ci,\\
\frac{dv}{dt} &= -bsv + ei - fv,
\end{aligned} \tag{52.7}$$

where, $a = \frac{\lambda K}{r},\ b = \frac{\gamma K}{r},\ c = \frac{\delta}{r},\ d = \frac{d_I}{r},\ e = \frac{\eta d_I}{r}$, and $f = \frac{\mu}{r}$.

52.2.1 Existence, Uniqueness and Boundedness

The right hand side of Eq. 52.7 are smooth functions of the variables s, i, v and parameters, as long as these quantities are non-negative, so local existence and uniqueness properties hold in the positive octant for some time interval $(0, t_f)$. In the next theorem we show that the linear combination of susceptible host, infected host and pathogens is less than a finite quantity or in other words, the solution of the system (52.7) is bounded and hence we get the theorem.

Theorem 52.1 *All the solution $y(t)$ of (52.7), where $y = (s, i, v)$, is uniformly bounded for $y_0 \in R^3_{0,+}$.*

52.2.2 Equilibrium Points and Their Stability Analysis

System (52.7) possesses the following equilibria: $E_0(0, 0, 0)$, $E_1(1, 0, 0)$, and $E^*(s^*, i^*, v^*)$, where,

$$i^* = \frac{(1-d-2s^*) + \sqrt{(1-d-2s^*)^2 + 4s^*(1-s^*)}}{2},$$
$$v^* = \frac{e(1-d-2s^*) + \sqrt{(1-d-2s^*)^2 + 4s^*(1-s^*)}}{2(bs^*+f)},$$

and s^* is the positive root of

$$\Omega_1 s^{*2} - \Omega_2 s^* - \Omega_3 = 0, \tag{52.8}$$

with $\Omega_1 = ab, \Omega_2 = bd + bc - af - be$ and $\Omega_3 = (d + c)f$.

Note that Eq. 52.8 has a unique positive root if $\Omega_1 > 0$, $\Omega_2 > 0$ and $\Omega_3 > 0$ for which $b(d + c) > af + be$, and i^* is positive if $\frac{1-d}{2} < s^* < 1$.

Now, to perform the system (52.7) is locally asymptotically stable or unstable for different equilibrium points, we get the following theorem:

Theorem 52.2 *The system (52.7) is unstable around E_0 for all parametric values.*

Theorem 52.3 *The system (52.7) is asymptotically stable around E_1 if $a + \frac{be}{b+f} < c + d$.*

Theorem 52.4 *The system (52.7) is stable around E^* if*

(i) $s^* + i^* > \frac{1}{2}$,

(ii) $\frac{abs^*}{e} < f < \frac{be}{a}$,

(iii) $\frac{b(d+c)}{b+f} > a$.

52.2.3 Global Stability for System (52.7)

The Equilibrium $E^*(s^*, i^*, v^*)$ is local asymptotic stable, we construct the Lyapunov function $U(s,i,v) = w_1(s - s^* - s^* \ln\frac{s}{s^*}) + w_2(i - i^* - i^* \ln\frac{i}{i^*}) + w_3(v - v^* - v^* \ln\frac{v}{v^*})$. Calculate the upper right derivation $U(s^*, i^*, v^*)$ along the system (52.7), we obtain

$$\begin{aligned} D^+U(s^*,i^*,v^*) &= w_1\frac{s-s^*}{s}\frac{ds}{dt} + w_2\frac{i-i^*}{i}\frac{di}{dt} + w_3\frac{v-v^*}{v}\frac{dv}{dt} \\ &= -w_1(s-s^*)^2 - 2w_2(s-s^*)(i-i^*) - b \\ &\quad (s-s^*)(v-v^*)(w_1+w_2) - a(w_1-w_2)(s-s^*) \\ &\quad (i-i^*) - w_1(s-s^*)\left(\frac{i^2}{s} - \frac{i^{*2}}{s^*}\right) + w_1(1+c) \\ &\quad (s-s^*)\left(\frac{i}{s} - \frac{i^*}{s^*}\right) + bw_2(i-i^*)\left(\frac{sv}{i} - \frac{s^*v^*}{i^*}\right) \\ &\quad + ew_3(v-v^*)\left(\frac{i}{v} - \frac{i^*}{v^*}\right) \\ &< 0. \end{aligned} \tag{52.9}$$

When, $w_1 \geq w_2$ and $s \in (0,\infty)$, $i \in (0,\infty)$ and $v \in (0,\infty)$, the minimum and maximum of $\frac{sv}{i}$, $\frac{i}{s}$ and $\frac{i}{v}$ are tends to zero respectively.

Thus, $D^+U(s^*, i^*, v^*) \leq 0$. According to the Lyapunov–LaSalle invariance principal, $E^*(s^*, i^*, v^*)$ is globally asymptotically stable.

52.3 Infective Class Giving Birth to Infective Offspring

Here we assuming to reframe the model system (52.6), that the infective class give birth to infective individual and grow in a logistic fashion. Also susceptible class S follows logistic growth with intrinsic growth rate "r_1" and carrying capacity "K_1". It is imperative that the infective class I also follows logistic growth with intrinsic growth rate "r_2" and carrying capacity "K_2".

With the above assumption the system (52.6) becomes

$$\begin{aligned} \frac{dS}{dt} &= r_1 S\left(1 - \frac{S}{K_1}\right) - \lambda SI - \gamma SV + \delta I, \\ \frac{dI}{dt} &= r_2 I\left(1 - \frac{I}{K_2}\right) + \lambda SI + \gamma SV - d_I I - \delta I, \\ \frac{dV}{dt} &= -\gamma SV + \eta d_I I - \mu V, \end{aligned} \tag{52.10}$$

where $S(0) > 0$, $I(0) > 0$, $V(0) > 0$.

The variables of model equations (52.10), are made dimensionless by rescaling all the variables in terms of carrying capacities. Thus we apply the transformation, $s = \frac{S}{K_1}, i = \frac{I}{K_2}, v = \frac{V}{K}, \tau = r_1 t, K = \frac{\lambda K_2}{\gamma}$ and get the following dimensionless form of the model equation (52.10). For notational convenience we will replace τ by t

$$\begin{aligned}
\frac{ds}{dt} &= s(1-s) - si - sv + a'i, \\
\frac{di}{dt} &= (e' - f' - g')i - e'i^2 + b'si + c'sv, \\
\frac{dv}{dt} &= -sv + mi - pv,
\end{aligned} \tag{52.11}$$

where, $a' = \frac{\delta K_2}{r_1 K_1}, b' = \frac{\lambda K_1}{r_1}, c' = \frac{K_1}{K_2}, e' = \frac{r_2}{r_1}, f' = \frac{d_I}{r_1}, g' = \frac{\delta}{r_1}, m = \frac{\eta d \gamma}{r_1 \lambda}$, and $p = \frac{\mu}{r_1}$.

Theorem 52.5 *All the solution $y'(t)$ of (52.11), where $y' = (s, i, v)$, is uniformly bounded for $y_0 \in (R')^3_{0,+}$*(Proof is obvious).

52.3.1 Equilibrium Points and Their Stability Analysis

System (52.11) possesses the following equilibria: $E'_0(0,0,0)$, $E'_1(1,0,0)$, and $E'^*(s^*, i^*, v^*)$. Here

$$i^* = \frac{(e' - g' - f')(s^* + p) + b's^*(s^* + p) + c'ms^*}{s^* + p}, \quad v^* = \frac{mi^*}{s^* + p},$$

and s^* is given by the positive root of the equation

$$\Omega'_1 (s^*)^4 + \Omega'_3 (s^*)^3 + \Omega'_2 (s^*)^2 + \Omega'_1 s^* - \Omega'_0 = 0, \tag{52.12}$$

with

$$\begin{aligned}
\Omega'_4 &= e(1 + b'), \\
\Omega'_3 &= 2e'p + e'(e' - g' - f') - e' + 2b'e'p + c'e'm + b'm + c'm^2, \\
\Omega'_2 &= e'p^2 + 2e'p(e' - g' - f' - 1 - a'b') + e'p^2(1 + b') + pm(b' + c'e') \\
&\quad + (e' - f' - g')(m - a'e') - a'c'e'm, \\
\Omega'_1 &= e'p^2(e' - g' - f' - 1 - a'b') - a'e'p(2e' - 2f' - 2g' + c'm) \\
&\quad + mp(e' - f' - g'),
\end{aligned}$$

and $\Omega'_0 = a'e'p^2(f' + g' - e')$.

Note that Eq. 52.12 has a unique positive root if $\Omega'_0 > 0$, $\Omega'_1 > 0$, $\Omega'_2 > 0$, $\Omega'_3 > 0$, and $\Omega'_4 > 0$ for which

(i) $f' + g' > e'$,

(ii) $p > \max\left(\frac{a'}{2}, 1, \frac{a'e'mp + 2pa(e' - f' - g') - m(e' - f' - g')}{1 - e' + g' + f' - a'b'}\right)$

(iii) $g' - a'b' > e' - f' - 1$.

Now, to check the stability of the system (52.11), we get the following theorem:

Theorem 52.6 *The system* (52.11) *is unstable around* E_0' *for all parametric values.*

Theorem 52.7 *The system* (52.11) *is asymptotically stable around* E_1' *if* $\frac{mc'}{p+1} + e' + b' < f' + g'$.

Theorem 52.8 *The system* (52.11) *is always stable around* E'^* *for all parametric values if*

(i) $i^* > \max\left(\frac{a'b'}{2e'}, \frac{a'b'}{e'}, \frac{c'v^*}{b'm}\right)$

(ii) $\max\left(\frac{a'b'}{c'}, \frac{e'i^*}{c'}, e'\right) < m < v^*e'$,

(iii) $s^* > \max\left(\frac{1 + a'b'}{b'}, a'\right)$,

(iv) $\max\left(\frac{a' + m}{a'm}, s^*\right) < c' < b'$.

It should be noted here that the system (52.11) is also globally stable around the interior equilibrium point.

52.4 Comparing Two Models with Numerical Simulations

Theoretical analysis of the model was done to explore stability, equilibria and uniqueness of the solutions and their boundedness and global stability for each system. But, for physical realization of the time evolution of different host and pathogen populations with varying model parameters, numerical solutions are considered for the set of Eqs. 52.6. Values of different constant model parameters, as given in Table 52.1, were chosen from the amassed literature in the field. Note that we want to emphasize the role of the recovery rate within the model. In Fig. 52.1 we have shown these hosts and pathogen populations as a function of time for $\delta = 0.2$. Note that, we have gone up to a time $t = 250$ days. This is because, a thorough check on the system reveals stabilization of all these populations well before $t = 250$ and within this the characteristic features of the system are manifested. In the left panel of Fig. 52.1 we find oscillatory solutions for all three populations bounded by stable upper and lower limits. As we increase δ, the upper and lower limits of solutions come closer (see the right panel of Fig. 52.1). Beyond some δ the two limits of solutions merge into one and thereafter, unique stable solutions for all three populations exist (see Fig. 52.2).

Table 52.1 Values of parameters used for models dynamics calculations

Parameters assigned	Definition	Default values (day^{-1})
r, r_1	Maximal growth rate of susceptible host	0.6, 0.8
r_2	Maximal growth rate of infected host	0.6
K	Carrying capacity	35 unit designated area
K_1	Carrying capacity of susceptible host	35 unit designated area
K_2	Carrying capacity of infected host	25 unit designated area
λ	Force of infection through contact with infected host	0.02 unit designated area
γ	Force of infection through contact with pathogens	0.04 unit designated area
d_I	Lysis death rate of infected host	2.5 l
η	Pathogens replication factor	115
μ	Mortality rate of pathogen	2.2

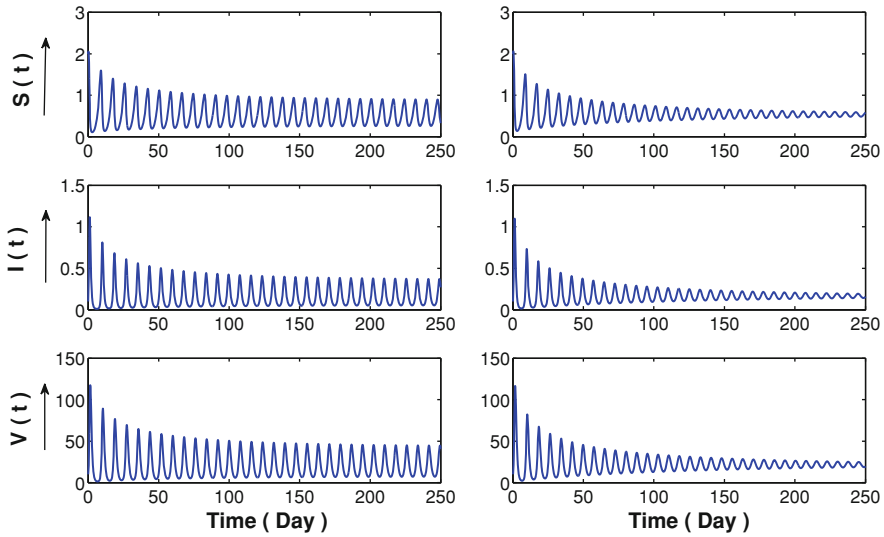


Fig. 52.1 Population densities of Susceptible host S, Infected host I and Pathogen V are plotted as a function of time for *left*: the Recovery rate $\delta = 0.2$; *right*: the recovery rate $\delta = 0.4$. Other parameter values are as in Table 52.1

Thus increase in the recovery rate δ reduces population fluctuation and bring stability to the system. Same results are observed for the system (52.10) with the parameter set given in Table 52.1 (for redundancy the results are not reported here).

To observe the disease transmission and it's effect on the system stability, we compare the two models through numerical simulation. Values of different constant model parameters for the system (52.6) and system (52.10) are taken from Table 52.1 (the common parameters have the same values). First we begin with the recovery rate within the model. For $\delta = 0.6$ the system (52.6) showed stability (see the left panel of Fig. 52.3) while system (52.10) showed oscillations (see the right

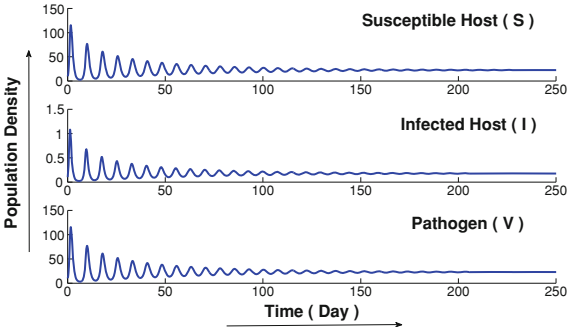


Fig. 52.2 Population densities of Susceptible host S, Infected host I and Pathogen V are plotted as a function of time for the Recovery rate $\delta = 0.6$. Other parameter values are as in Table 52.1

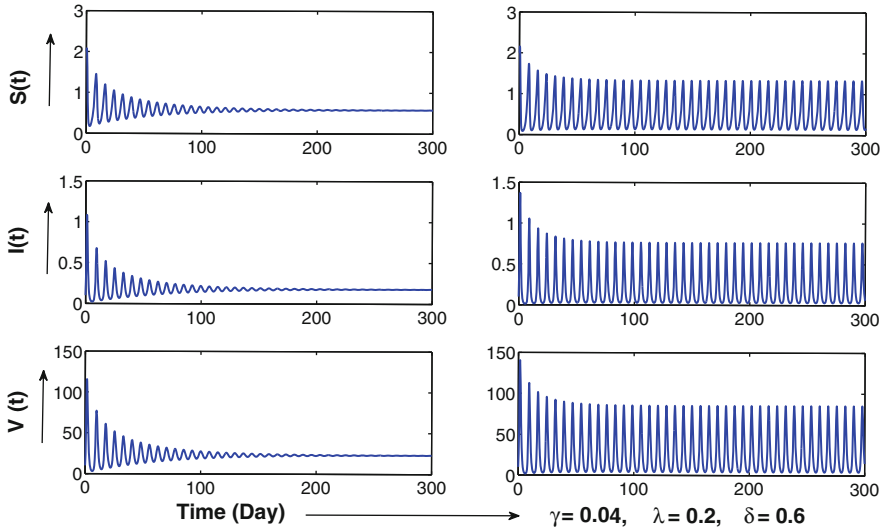


Fig. 52.3 *Left*: the system (52.6) with recovery rate $\delta = 0.6$ and other parameter values are given in Table 52.1; *right*: the system (52.10) with $\delta = 0.6$ and other parameter values are given in Table 52.1

panel of Fig. 52.3). Further comparing the spreading rate of the disease, it is observed that for the system (52.6) the system showed stability even in case of high individual contact rate while the system (52.10) showed oscillations, see Fig. 52.4. When it comes to pathogen contact rate, we observed that small amount of pathogen contact can also destabilizes the system (52.10), see Fig. 52.5.

52.5 Discussion

To start discussion we first recall our main aim of the present study—whether the vertical transmission increase virulence or not? To find way to answer this question we proposed and analyzed two model systems. In the system (52.6) we

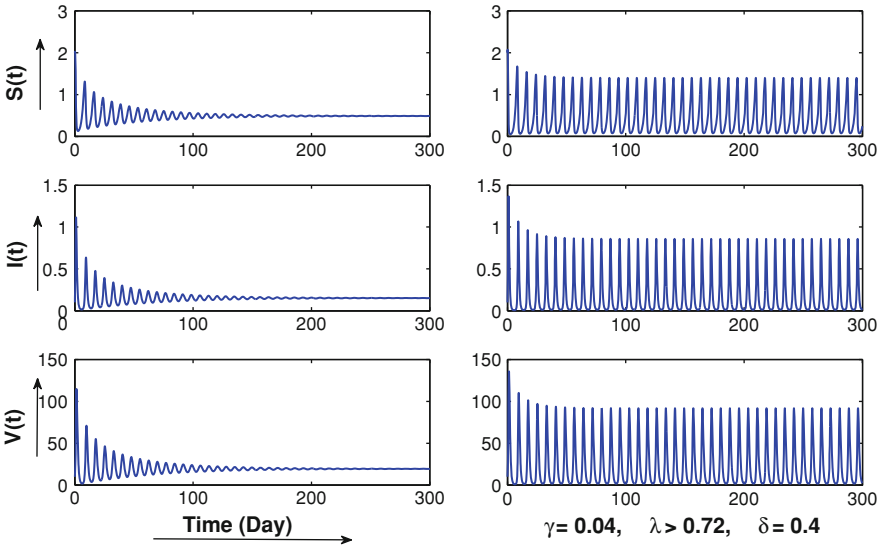


Fig. 52.4 *Left*: the system (52.6) with force of infection through individual contact $\lambda = 0.72$ and other parameter values are given in Table 52.1; *right*: the system (52.10) with $\lambda = 0.72$ and other parameter values are given in Table 52.1

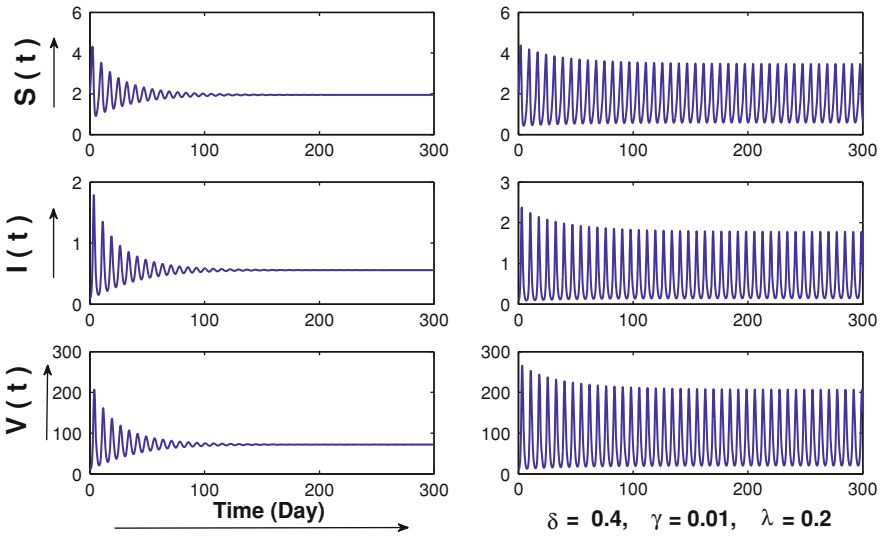


Fig. 52.5 *Left*: the system (52.6) with force of infection through pathogen contact $\gamma = 0.01$ and other parameter values are given in Table 52.1; *right*: the system (52.10) with $\gamma = 0.01$ and other parameter values are given in Table 52.1

considered the transmission to be purely horizontal (though the infected population still reproduce) and in the system (52.10) the transmission is considered to be mixed, i.e., both horizontal and vertical transmissions are present in the system. We separately analyzed the two systems. We observed that in both the system the solutions exists and are unique and bounded. The equilibrium point and their stability analysis showed that both the system have same kind of equilibrium point and their stability natures are same but the parametric conditions are different. To explore parameter dependence of each equilibrium point we performed the numerical simulations.

Global stability are also observed for both the systems and it showed that both the systems are globally stable around the interior equilibrium point under certain parametric conditions.

Our numerical analysis showed that the recovery rate plays a vital role in maintaining the stability of the system. This result is independent of the presence or absence of the vertical transmission. Further analysis showed that in case where the vertical transmission is present it becomes difficult to control the fluctuation of the system population. It was observed that in case of disease with both vertical and horizontal transmission, a small pathogenic contact can also destabilizes the system. In this case even high recovery rate could not be enough to control the fluctuation and maintain the stability of the system. Thus answering our question, we could say that the presence of vertical transmission along with the horizontal transmission increases the virulence of the pathogens.

Acknowledgments Research is supported by the Department of Mathematics, Jadavpur University, PURSE DST, Government of India.

References

1. Mansfield BE, Dionne MS, Schneider DS, Freitag NE (2003) Exploration of host–pathogen interactions using *Listeria monocytogenes* and *Drosophila melanogaster*. Cell Microbiol 5:901–911
2. Ross R (1916) An application of theory of probabilities to the study of a priori pathometry. Part I. Proc R Soc Lond Ser A 92:204–230
3. Ross R, Hudson HP (1917) An application of theory of probabilities to the study of a priori pathometry. Part II and part III. Proc R Soc Lond Ser A 93:212–240
4. Roy PK, Chattopadhyay B (2010) Host pathogen interactions with recovery rate: a mathematical study. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, WCE 2010, 30 June–2 July, London, UK, pp 521–526
5. Anderson RM, May RM (1981) The population dynamics of microparasites and their invertebrate hosts. Philos Trans R Soc Lond B 291:451–524
6. Anderson RM, May RM (1991) Infectious diseases of humans, dynamics and control. Oxford Science Publications, Oxford
7. Begon M, Bowers RG (1995) Host–host pathogen models and microbial pest control: the effect of host self regulation. J Theor Biol 169:275–287

8. Holt RD, Pickering J (1985) Infectious disease and species coexistence: a model in Lotka–Volterra form. Am Nat 126:196–211
9. Kribs-Zaleta CM, Velasco-Hernandez JX (2000) A simple vaccination model with multiple endemic states. Math Biosci 164(2):183–201
10. Moghadas SM (2004) Modelling the effect of imperfect vaccines on disease epidemiology. Discrete Continuous Dyn Syst Ser B 4(4):999–1012
11. Brankston G, Gitterman L, Hirji Z, Lemieux C, Gardam M (2007) Transmission of influenza A in human beings. Lancet Infect Dis 7:257–265
12. Attye A, Auger P, Joly J (1990) Incidence of occult athlete's foot in swimmers. Eur J Epidemiol 6:244–247
13. De Cock KM et al (2000) Prevention of mother-to-child HIV transmission in resource-poor countries. J Am Med Assoc 283:1175–1182
14. Woo D, Cummins M, Davies PA, Harvey DR, Hurley R, Waterson AP (1979) Vertical transmission of hepatitis B surface antigen in carrier mothers in two west London hospitals. Arch Dis Child 54:670–675
15. Favre M, Majewski S, De Jesus N, Malejczyk M, Orth G, Jablonska S (1998) A possible vertical transmission of human papillomavirus genotypes associated with epidermodysplasia verruciformis. J Invest Dermatol 111:333–336
16. McCallum H, Barlow N, Hone J (2001) How should pathogen transmission be modelled? Trends Ecol Evol 16:295–300
17. Lipsitch M, Siller S, Nowak MA (1996) The evolution of virulence in pathogens with vertical and horizontal transmission. Evolution 50:1729–1741
18. Bull JJ, Molineux I, Rice WR (1991) Selection of benevolence in a host–parasite system. Evolution 45:875–882
19. Ewald PW (1987) Transmission modes and evolution of the parasitism–mutualism continuum. Ann NY Acad Sci 503:295–306
20. De Jong MCM, Diekmann O, Heesterbeek JAP (1994) How does infection transmission depend on population size? In: Mollison D (ed) Epidemic models, their structure and relation in data. Cambridge University Press, Cambridge

Chapter 53
Chaotic Oscillations in Hodgkin–Huxley Neural Dynamics

Stimulus Reconstruction and Neural Dynamics Retrieval

Mayur Sarangdhar and Chandrasekhar Kambhampati

Abstract Neural responses are the fundamental expressions of any neural activity. Information carried by a neural response is determined by the nature of a neural activity. In majority of cases the underlying stimulus that triggers it remains largely unknown. Previous studies to reconstruct the stimulus from a neural response show that the high non-linearity of neural dynamics renders inversion of a neuron a challenging task. This paper presents a numerical solution rather than an analytical one to reconstruct stimuli from Hodgkin–Huxley neural responses. The stimulus is reconstructed by first retrieving the maximal conductances of the ionic channels and then solving the Hodgkin–Huxley equations for the stimulus. The results show that the reconstructed stimulus matches the original stimulus to a high degree of accuracy. In addition, this reconstruction approach also retrieves the neural dynamics for which an analytical solution does not currently exist. Constant-current and periodic stimuli are shown to be accurately reconstructed using this approach.

53.1 Introduction

The encoding and decoding mechanisms adopted by neurons can be understood by studying the relationship between a neural response and its stimulus. How neurons specifically encode and decode information remains a challenging question as not much is known about neural coding. It is thought that either the firing time or the

M. Sarangdhar (✉) · C. Kambhampati
Department of Computer Science, University of Hull, Cottingham Road, East-Yorkshire, Hull, HU6 7RX, UK
e-mail: M.Sarangdhar@2006.hull.ac.uk

C. Kambhampati
e-mail: C.Kambhampati@hull.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_53,
© Springer Science+Business Media B.V. 2011

rate of fire of a neuron carries specific neural response information [1–3]. Some research suggests that reconstructing the stimulus from a neural response can help understand neural processing. A stimulus represents a trigger for a neural activity which underlines any neural response. The ability to reconstruct a stimulus hence offers to retrieve stimulus parameters that can help extend our understanding of neuronal encoding/decoding.

Previous work on input reconstruction has been carried out across many fields like digital filters, neural networks, algorithms and complexity, and digital signal processing [4–13]. Similar approach can be considered for stimulus reconstruction; however, due to the high non-linearity of neural dynamics, it is very difficult to obtain an analytic solution. Periodic signals, unlike aperiodic signals, can be recovered using conventional filters [4]. Artificial neural networks are used to treat the Hodgkin–Huxley (HH) neuron as a black box and reconstruct the stimulus by learning the dynamics [5]. Other implementations use a reverse filter that predicts the sensory input from neuronal activity and recursive algorithms to reconstruct stimulus from an ensemble of neurons [6, 7]. The principles of a time encoding and decoding machines for signal recovery have been explored to reconstruct a neural stimulus whereas, a more direct approach to recover stimulus focuses to make the HH neuron input–output (IO) equivalent to an integrate and fire (IF) neuron [8–13]. These approaches establish a relationship between the neural response and the stimulus but are not designed to capture or retrieve the neural dynamics. In other words, they offer some starting point for stimulus reconstruction but it is quite a challenge to analytically invert a neuron. However, it is possible to reconstruct stimulus from a neural response using numerical approximations and small time-steps for integration.

This demonstrates the reconstruction of constant-current and periodic stimuli by (a) extracting the maximal conductances from a trace of neural response and (b) solving the neural equations for the stimulus. To reconstruct the stimulus, it is imperative that linearization is carried out. This chapter demonstrates the above approach using a Hodgkin–Huxley (HH) neuron [14] and Euler integration. The results show that for a small time-step δ, the accuracy of extracted maximal conductances is very high. Also, the reconstructed stimulus matches the original stimulus accurately. As reconstruction of the stimulus involves solving the neural equations, this approach can replicate the neural dynamics, the time-dependent changes in the voltage-gated ionic channels of Na^+, K^+ and Cl^-. This technique, though computationally demanding, offers a local solution to the problem of inverting a neural response.

53.2 Neuronal Model and Synapse

53.2.1 The Neuron Model

The computational model and stimulus for an HH neuron is replicated from [15]. The differential equations of the model are the result of non-linear interactions between the membrane voltage V and the gating variables m, h and n for Na^+, $K^+$and Cl^-.

53.2.2 The Synaptic Current

An input spike train given by Park and Kim [16] is considered to generate the pulse component of the external current.

$$I_P = g_{\text{syn}} \sum_n \alpha(t - t_f)(V_a - V_{\text{syn}}) \tag{53.1}$$

g_{syn}, V_{syn} are the conductance and reversal potential of the synapse. The α-function is defined in [16] as

$$\alpha(t) = (t/\tau)e^{-t/\tau}\Theta(t), \tag{53.2}$$

where τ is the time constant of the synapse and $\Theta(t)$ is the Heaviside step function. $V = 30$ mV, $\tau = 2$ ms, $g_{\text{syn}} = 0.5$ ms/cm^2 and $V_{\text{syn}} = -50$ mV.

53.2.3 The Total External Current

The total external current applied to the neuron is a combination of static and pulse component

$$I_i = I_S + I_P + \varepsilon \tag{53.3}$$

where I_S is the static and I_P is the pulse current, ε is the random Gaussian noise with zero mean and standard deviation $\sigma = 0.025$. On injection of a periodic or sinusoidal stimulus the steady state response of a neuron is no longer preserved [17–25]. The self-excited oscillations of the HH neuron [14] may become chaotic when a sinusoidal stimulus is applied with proper choices of magnitude and frequency [20, 21, 25, 26]. Physiological experiments on squid giant axons [18, 19] and Onchidium neurons [22] have confirmed the occurrence of chaotic oscillations. It is understood that distinct sinusoidal stimuli induce different chaotic oscillations which result in dissimilar neural responses [27–29].

53.3 Stimulus Reconstruction

Let $V(t)$ be the neural response of the HH neuron to a synaptic stimulus $I(t)$ and ionic conductances g_{Na}, g_K and g_L. Assuming that $I(t)$ is unknown and only the neural response and the reversal potentials are known, the aim is to reconstruct the stimulus $I'(t)$ such that $I(t)$ and $I'(t)$ are identical. Therefore the target is to retrieve g'_{Na}, g'_K and g'_L and get $I'(t)$ without any information of $I(t)$.

53.3.1 Extracting Maximal Conductances

The gating variables m, n and h only depend on the instantaneous voltage at time t. The instantaneous voltage at time t is given by

$$v(t) = v(0) + \frac{1}{C}\begin{bmatrix} \int_0^t I(t')dt' - g'_{\mathrm{Na}} \int_0^t m^3(t')h(t') \cdot (v(t') - V_{\mathrm{Na}})dt' \\ -g'_K \int_0^t n^4(t') \cdot (v(t') - V_K)dt' - g_L \int_0^t (v(t') - V_L)dt' \end{bmatrix} \tag{53.4}$$

To retrieve the three ionic conductances, linear equations in three unknowns need to be solved. The formulation of the equations is proposed as an algorithm in [30]. Given a small voltage trace $v(t)$, select three times ti, $i = 1, 2, 3$. As the voltage trace $v(t)$ is known over all t, $v(t)$ is known for $i = 1, 2, 3$.

Let functions $f_j(t)$, $j = 1, 2, 3$ be defined as

$$\begin{aligned} f_1(t) &= -\frac{1}{C}\int_0^t m^3(t')h(t') \cdot (v(t') - V_{Na})dt' \\ f_2(t) &= -\frac{1}{C}\int_0^t n^4(t') \cdot (v(t') - V_K)dt' \\ f_3(t) &= -\frac{1}{C}\int_0^t (v(t') - V_L)dt' \end{aligned} \tag{53.5}$$

and $b(t)$ defined as

$$b(t) = v(t) - v(0) - \int_0^t I(t')dt' \tag{53.6}$$

Hence,

$$b(t) = g'_{\mathrm{Na}} f_1(t) + g'_K f_2(t) + g'_L f_3(t) \tag{53.7}$$

If $\int_0^t I(t')dt'$ is a known analytic function, the value of $b(t)$ is known for all values of t. Hence, for a voltage trace $v(t)$ and external stimulus $I(t)$, approximations to the gating variables, m, n and h are obtained by integrating the HH equations. If m' n' and h' are the gating-variables' estimates and $f'_j(t)$ is the

resultant approximation of $f_j(t)$, then the retrieving maximal conductances can be defined as a solution to the linear system

$$b(t_i) = \sum_{j=1}^{3} f_j^{'}(t_i)x_j, \quad i = 1, \ldots, N \tag{53.8}$$

This is an overdetermined system of linear equations in the form $Ax = b$. An approximate solution can be obtained by using the full set of data generated during the integration of the HH equations and treating (53.8) as a linear least squares problem.

Hence, the best fit solution in the linear least squares sense is obtained by solving

$$\min_x \sum_{i=1}^{N} \left(b(t_i) - \sum_{j=1}^{3} f_j'(t_i)x_j \right)^2 \tag{53.9}$$

If $A^\delta \in \mathbb{R}^{\mathbb{N}\times\not\Vdash}$ is the matrix whose entries are $a_{i,j}^\delta = f_j^{'}(t_i), i = 1, \ldots, N$ and $b \in \mathbb{R}^{\mathbb{N}}$,

$$\min_x \left\| A^\delta x - b \right\|_2 \tag{53.10}$$

As the equations $Ax = b$ are linear in x, a solution is obtainable.

53.3.2 Reconstructing the Stimulus

The approach defined above requires the knowledge of both the voltage $v(t)$ and the external stimulus $I(t)$, for all time t. In principle, it is unrealistic to know the stimulus for all times t and in majority cases, the stimulus $I(t)$ remains unknown. Therefore, retrieving the maximal conductances using Eq. 53.11 is specific when all parameters are known.

However, it is possible to reconstruct the stimulus entirely without the knowledge of corresponding $I(t)$for a neural response $V(t)$. As the type of the neuron and the reversal potential for Na^+, K^+ and Cl^- is known, we propose that the neural stimulus can be reconstructed without the knowledge of the original stimulus $I(t)$.

1. Record any neural response $V(t)$ whose stimulus, say $I(t)$, requires to be reconstructed
2. Inject a supra-threshold stimulus, $I_s(t_s)$ for a small time duration t_s
3. Record the corresponding voltage trace generated, $v_s(t_s)$
4. Retrieve the maximal conductances using Eq. 53.11 and $I_s(t_s)$ as the external stimulus

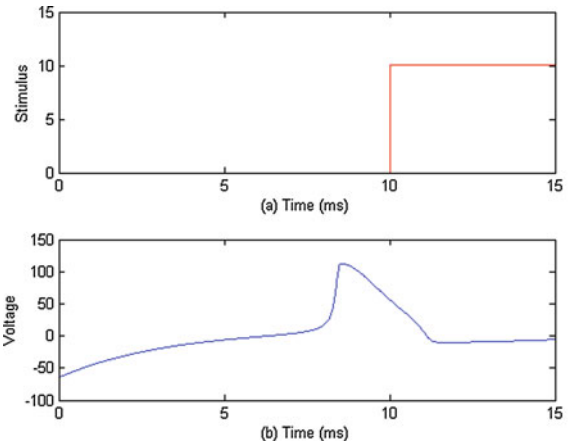


Fig. 53.1 The voltage trace v_s generated by a small step-current I_s. This small trace of neural voltage is sufficient to retrieve the maximal conductances

5. Using the approximated maximal conductances, g'_{Na}, g'_K and g'_L, solve the HH equations using the recorded neural response $V(t)$ and the stimulus as the only unknown to get the reconstructed stimulus $I'(t)$

   The HH equations can be re-written as

$$\left.\begin{aligned} I'(t) = g'_{\mathrm{Na}} m'(t)^3 h'(t)(v(t) - V_{\mathrm{Na}}) + g'_K n'(t)^4 (v(t) - V_K) \\ + g'_L (v(t) - V_L) + C\frac{dv}{dt} \end{aligned}\right\} \tag{53.11}$$

where g'_{Na}, g'_K and g'_L are the approximated maximal conductances calculated from $v_s(t_s)$ and m', n' and h' are the estimates of the gating variables m, n and h respectively.

As $V(t)$ is known for all times t, the rate of change of voltage (dv/dt) can be numerically approximated.

This approach provides a local solution to reconstructing the neural stimulus of a HH neuron and also approximates the gating variables. In addition to the retrieval of stimulus parameters, it also estimates the neural dynamics which are important represent the open-close mechanism of ionic gates.

53.4 Computational Results

53.4.1 Generating a Voltage Trace

Let I_s be a small supra-threshold step current that evokes an action potential. The resultant voltage trace v_s is sufficient to retrieve the maximal conductance values (Fig. 53.1).

Table 53.1 Retrieved maximal conductance values for various values of δ

Original↓/Retrieved→	$\delta = 0.01$	$\delta = 0.001$	$\delta = 0.0001$
$g_{Na} = 120$	$g'_{Na} = 120.49$	$g'_{Na} = 120.05$	$g'_{Na} = 120$
$g_K = 36$	$g'_K = 36$	$g'_K = 36$	$g'_K = 36$
$g_L = 0.30$	$g'_L = 0.33$	$g'_L = 0.30$	$g'_L = 0.30$

The conductances are highly accurate as δ becomes close to 0

53.4.2 Retrieving Maximal Conductances

Given the voltage trace v_s and the corresponding external stimulus I_s, near approximation of the maximal conductance values can be obtained using Eq. 53.11. Let δ be the time-step of the Euler integration. It is observed that the accuracy of the approximated conductances is dependent on δ. Accuracy increases if δ chosen is close to 0. These approximated conductances are consistent with the observations of [30]. As (15) is an overdetermined system of linear equations, an exact solution cannot be obtained for all values of δ (Table 53.1).

The relative error of the approximations decreases as δ becomes close to 0 (Table 53.2).

The voltage traces reconstructed from the approximated conductances are shown in Fig. 53.2. The estimated maximal conductance values produce a good fit to the original trace v_s.

53.4.3 Stimulus Reconstruction

The retrieval of maximal conductance values such that a good fit of the original voltage trace is produced indicates that the approximations are nearly accurate.

53.4.3.1 Constant-Current Stimulus

Let the HH neuron be stimulated by an unknown step-current I_{step} such that it evokes a series of action potentials V_{step}. The maximal conductances are approximated in Table 53.1. The reconstructed stimulus is shown in Fig. 53.3.

Results show that if the time-step of Euler integration is sufficiently small i.e. $\delta \sim 0.0001$, the maximal conductances can be accurately retrieved. The stimulus reconstructed using these maximal conductance values, is a near approximation of the original unknown stimulus.

53.4.3.2 Periodic Stimulus

If the HH neuron is stimulated by an unknown periodic stimulus I_{periodic}, the resultant neural response is V_{periodic}.

Table 53.2 The relative error ε decreases as δ becomes close to 0

δ	Relative error (ε)
0.01	0.0037
0.001	0.00038
0.0001	0

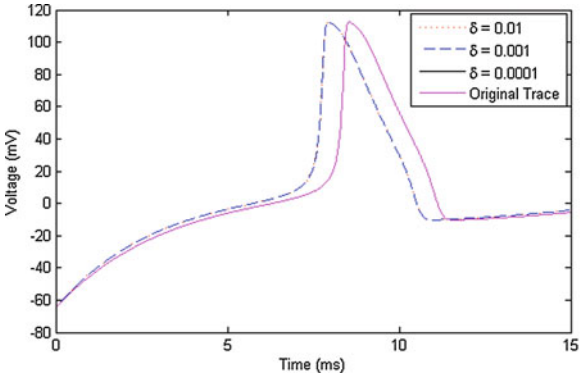
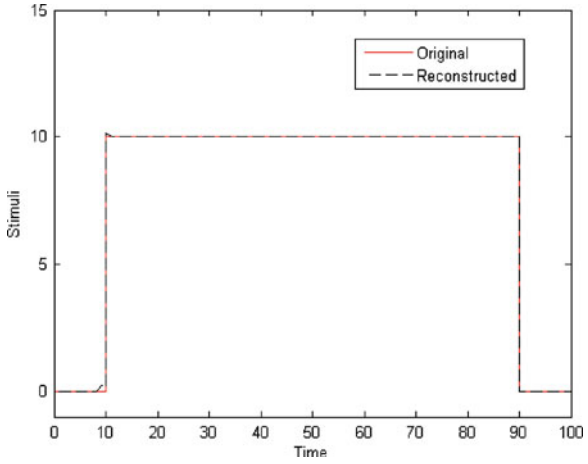


Fig. 53.2 The reconstructed voltage trace using the approximated maximal conductance values for different time-stepsδ. As δ becomes close to 0, the approximations approach the actual conductance values. For $\delta = 0.0001$, the approximated conductance values are equal to the original values. Hence the trace generated by $\delta = 0.0001$ overlaps with the original trace v_s

Fig. 53.3 The reconstructed stimulus is good fit to the original stimulus. The original stimulus is very well approximated if chosen δ is close to 0



It is observed that the unknown stimulus can be predicted accurately if δ is small and close to 0. As a result, the computational time required by this approach is directly proportional to the choice of δ. However, this is approach provides a local solution to reconstructing unknown stimuli using the knowledge of the computational model of a neuron. It is also possible to retrieve the neural dynamics which cannot be retrieved by a purely analytical approach (Fig. 53.7).

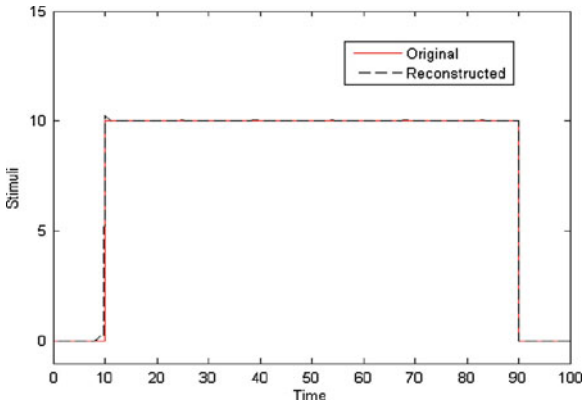


Fig. 53.4 The approximations become less accurate with an increase in δ

53.5 Conclusions

The neural dynamics of the HH neuron have been the subject of research for many years now. The dynamics put forth by Hodgkin and Huxley have been well studied and replicated by many researchers. In much the same way, inverting the HH neural equations has attracted interest in recent years. The equations of the HH neuron are highly non-linear due to the incorporation of probability of the gating variables m, n and h which regulate the open-close mechanism of ionic channels.

variables m, n and h and their time constants τ_m, τ_n and τ_h.

Previous research has addressed the problem of inverting this non-linear neuron by using digital filters, neural networks, algorithms and complexity, and digital signal processing. Other approaches point to the use of reconstruction algorithms, time encoding/decoding machines or an IF neuron. These approaches establish a relationship between the neural response and the stimulus but they are not designed to capture or retrieve the neural dynamics (Figs. 53.4 and 53.5).

The approach described in this chapter provides a numerical solution to reconstruct an unknown neural stimulus. Using the approximated maximal conductances, g'_{Na}, g'_K and g'_L, solve the HH equations using the recorded neural response $V(t)$ and the stimulus as the only unknown to get the reconstructed stimulus $I'(t)$(Figs. 53.6 and 53.7).

It is observed that the accuracy of maximal conductances retrieved by solving an overdetermined system of linear equations depends on the time-step (δ) of Euler integration. A small value of $\delta \sim 0.0001$ can reproduce almost exact maximal conductances. Accurate maximal conductance values help reconstruct a near-fit approximation of the original stimulus. Due to the nature of numerical approximation and the inherent non-linearity in the neural dynamics, the reconstructed stimulus shows some jitters. Also, it is noticed that if the original stimulus carries any noise, an exact match of the stimulus cannot be reconstructed. However, the reconstructed stimulus still matches the original stimulus to a high degree of accuracy. The choice of δ is very important and there is a trade-off between computational time and accuracy. The accuracy increases with a decrease in δ.

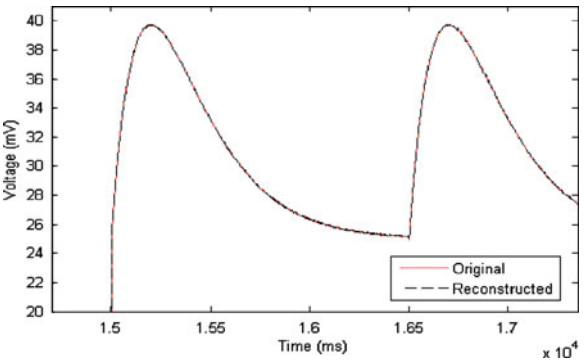


Fig. 53.5 The reconstructed periodic stimulus for δ close to 0. For $\delta = 0.0001$, the reconstructed stimulus is a near-fit of the original stimulus

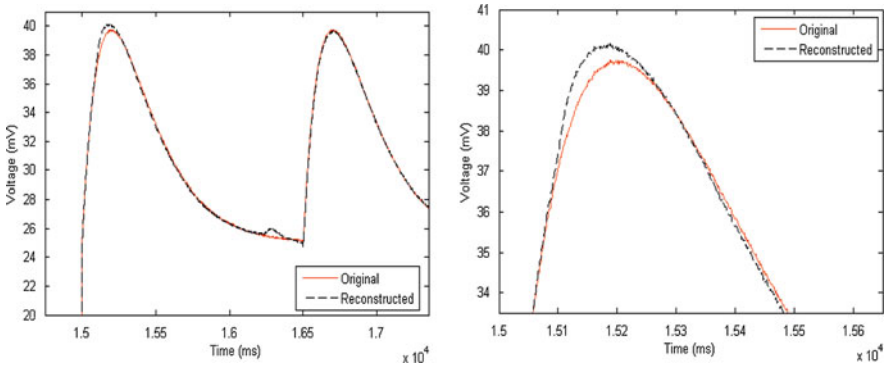


Fig. 53.6 The approximation of the reconstructed stimulus become less accurate with an increase in δ ($\delta \sim 0.001$). The numerical approximation of the derivatives causes some jitters. The jitters are due to the numerical approximation to the rate of change of voltage. However, the reconstruction is very close to the original stimulus for δ close to 0

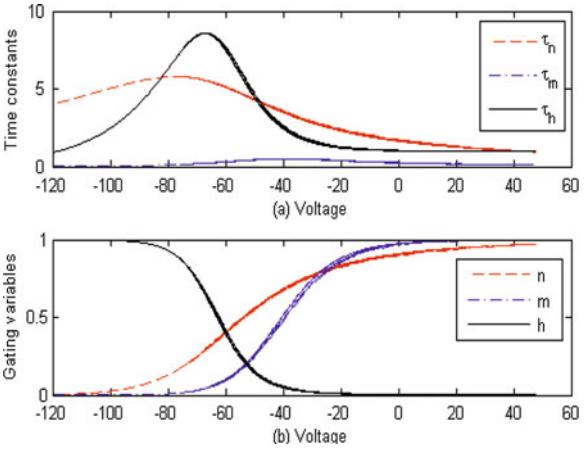


Fig. 53.7 The reconstructed neural dynamics. This numerical solution can retrieve the gating potentials of the neuron during an action potential

The approached described in this chapter can reconstruct very good approximations of the original stimuli. The results show that the unknown periodic and constant current stimuli are well approximated by this reconstruction method. It is also worth mentioning that although establishing an IO relationship can provide some information of the stimulus parameters, the current approach can accurately reconstruct the neural dynamics in addition to an unknown stimulus.

References

1. Rinzel J (1985) Excitation dynamics: insights from simplified membrane models. Theor Trends Neurosci Fed Proc 44(15):2944–2946
2. Panzeri S, Schultz SR, Treves A, Rolls ET (1999) Correlations and the encoding of information in the nervous system. Proc R Soc Lond B 266:1001–1012
3. Gabbiani F, Metzner W (1999) Encoding and processing of sensory information in neuronal spike trains. J Biol 202:1267–1279
4. Das A, Folland R, Stocks NG, Hines EL (2006) Stimulus reconstruction from neural spike trains: are conventional filters suitable for both periodic and aperiodic stimuli? Signal Process 86(7):1720–1727
5. Saggar M, Mericli T, Andoni S, Miikkulainen R (2007) System identification for the Hodgkin–Huxley model using artificial neural networks. IEEE international joint conference on neural networks orlando, FL, pp 2239–2244, 12–17 Aug 2007
6. Stanley GB, SeyedBoloori A (2001) Decoding in neural systems: stimulus reconstruction from nonlinear encoding. In: Proceedings of the 23rd annual international conference of the IEEE engineering in medicine and biology society, vols 1–4, 23. pp 816–819
7. Stanley GB (2001) Recursive stimulus reconstruction algorithms for real-time implementation in neural ensembles. Neurocomputing 38:1703–1708
8. Lazar AA, Pnevmatikakis EA (2009) Reconstruction of sensory stimuli encoded with integrate-and-fire neurons with random thresholds. EURASIP Journal on advances in Signal Processing. Article no. 682930
9. Lazar AA (2007) Information representation with an ensemble of Hodgkin–Huxley neurons. Neurocomputing 70(10–12):1764–1771
10. Lazar AA (2007) Recovery of stimuli encoded with Hodgkin–Huxley neurons. Computational and systems neuroscience meeting, COSYNE 2007, Salt Lake City, UT, February 22–25
11. Lazar AA (2006) Time encoding machines with multiplicative coupling, feedforward, and feedback. IEEE Trans Circuits Syst II Express Briefs 53(8):672–676
12. Lazar AA (2004) Time encoding with an integrate-and-fire neuron with a refractory period. Neurocomputing 58:53–58
13. Lazar AA, Simonyi EK, Toth LT (2006) A real-time algorithm for time decoding machines. 14th European signal processing conference, Sept 2006
14. Hodgkin A, Huxley A (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 117:500–544
15. Hasegawa H (2000) Responses of a Hodgkin–Huxley neuron to various types of spike-train inputs. Phys Rev E 61(1):718–726
16. Park MH, Kim S (1996) Analysis of phase models for two coupled Hodgkin–Huxley neurons. J Kr Phys Soc 29(1):9–16
17. Wang XJ, Buzsáki G (1996) Gamma oscillation by synaptic inhibition in a hippocampal interneuronal network model. J Neurosci 16(2):6402–6413
18. Guttman R, Feldman L, Jakobsson E (1980) Frequency entrainment of squid axon membrane. J Membr Biol 56:9–18

19. Matsumoto G, Kim K, Ueda T, Shimada J (1980) Electrical and computer simulations upon the nervous activities of squid giant axons at and around the state of spontaneous repetitive firing of action potentials. J Phys Soc Jpn 49:906
20. Aihara K, Matsumoto G, Ikegaya Y (1984) Periodic and non-periodic responses of a periodically forced Hodgkin–Huxley oscillator. J Theor Biol 109:249–269
21. Matsumoto G, Aihara K, Ichikawa M, Tasaki A (1984) Periodic and nonperiodic responses of membrane potentials in squid giant axons during sinusoidal current simulations. J Theor Neurobiol 3:1–14
22. Hayashi h, Ishizuka S, Hirakawa K (1985) Chaotic responses of the pacemaker neuron. J Phys Soc Jpn 54:2337
23. Holden AV (1987) Chaos. Manchester University Press, Manchester
24. Hasegawa H (2008) Information conveyed by neuron populations—firing rate, fluctuations and synchrony. Neuroquantology 6(2):105–118
25. Kaplan D, Glass L (1995) Understanding non linear dynamics. Springer, New York
26. Wilson HR (1999) Spikes, decisions and actions—the dynamical foundations of neuroscience. Oxford University Press Inc., New York
27. Sarangdhar M, Kambhampati C (2008) Spiking neurons: is coincidence-factor enough to compare responses with fluctuating membrane voltage? In: world congress on engineering 2008: the 2008 international conference of systems biology and bioengineering, vol 2. pp 1640–1645, London, UK, 2–4 July 2008
28. Sarangdhar M, Kambhampati C (2008) Spiking neurons and synaptic stimuli: determining the fidelity of coincidence-factor in neural response comparison. Eng Lett 16(4):512–517
29. Sarangdhar M, Kambhampati C (2009) Spiking neurons and synaptic stimuli—neural response comparison using coincidence-factor. In: Gelman L, Balkan N, Ao S (eds) Advances in electrical engineering, computational science. Springer, The Netherlands
30. Shepardson D (2009) Algorithms for inverting Hodgkin–Huxley type neuron models. PhD dissertation, Georgia Institute of Technology. http://www.aco.gatech.edu/doc/Shepardson_thesis.pdf. Accessed Nov 2009

Chapter 54
Quantification of Similarity Using Amplitudes and Firing Times of a Hodgkin–Huxley Neural Response

Mayur Sarangdhar and Chandrasekhar Kambhampati

Abstract Periodic stimuli are known to induce chaotic oscillations in the squid giant axon for a certain range of frequencies. This behaviour is modelled by the Hodgkin–Huxley equations when a periodic stimulus is applied. The responses resulting from chaotic neural dynamics have irregular inter-spike intervals and fluctuating amplitudes. These characteristics are absent in steady state responses of the Hodgkin–Huxley neuron which are generated using a constant current stimulus. It is known that firing time information is adequate to estimate similarity of steady state responses; however, in the presence of chaotic oscillations, similarity between neural responses cannot be estimated using firing time alone. The results discussed in this paper present a quantification of the similarity of neural responses exhibiting chaotic oscillations by using both amplitude fluctuations and firing times. We relate the similarity thus obtained between two neural responses to their respective stimuli. Identical stimuli have very similar effect on the neural dynamics and therefore, as the temporal inputs to the neuron are the same, the occurrence of identical chaotic patterns result in a high estimate of similarity for the neural responses. Estimates of similarity are compared for periodic stimuli with a range of inter-spike intervals.

M. Sarangdhar (✉) · C. Kambhampati
Department of Computer Science, University of Hull, Cottingham Road, Hull, East-Yorkshire HU6 7RX, UK
e-mail: M.Sarangdhar@2006.hull.ac.uk

C. Kambhampati
e-mail: C.Kambhampati@hull.ac.uk

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_54,
© Springer Science+Business Media B.V. 2011

54.1 Introduction

The non-linear dynamics of a neuron have been studied both theoretically and physiologically in recent years to extend the understanding of its underlying mechanism [1–13]. The spikes or action potentials are evoked when an external stimulus is applied to the neuron. It is thought that either the firing rate or firing time of individual spikes carries specific information of the neuronal response [14–16]. This holds for all steady state responses of a neuron when a constant current stimulus is applied. However, on injection of a periodic or sinusoidal stimulus the steady state response is no longer preserved [17–26]. The self-excited oscillations of a Hodgkin–Huxley (HH) neuron [27] may become chaotic when a sinusoidal stimulus is applied with proper choices of magnitude and frequency [20, 21, 25, 26]. Physiological experiments on squid giant axons [18, 19] and Onchidium neurons [22] have confirmed the occurrence of chaotic oscillations. This paper quantifies the similarity estimated between neural responses exhibiting chaotic oscillations. By using the amplitude distribution and the firing times of a neural spike train to estimate similarity.

The nature of a periodic stimulus is responsible to induce chaotic oscillations in a biological neuron. Irregular inter-spike interval (ISI) and fluctuating amplitudes are the characteristics of chaotic oscillations absent in steady state responses generated by constant current stimuli. Information on stimuli similarity can be derived from neural response comparison. Firing time information is adequate to estimate similarity of steady state responses; however, in the presence of chaotic oscillations or when the amplitudes of a neural response fluctuate, amplitude and firing time collectively reflect the true dynamics of a neuron and therefore both should feature in similarity estimation [28–30]. Similarity estimation is based on the principle of relative coincidences without coincidences by chance [31, 32]. The amplitudes a neural response exhibiting these chaotic oscillations fit a Normal distribution and it is possible to determine amplitude coincidences using the properties of Normal distribution. Similarity between these responses can be estimated by a composite similarity measure based on amplitude and firing time coincidences. Results show that similarity based on this composite approach is mathematically realisable than similarity based on firing times or amplitudes alone. It is observed that similar periodic stimuli induce similar chaotic patterns in the neural responses and therefore the resulting neural responses have a high degree of similarity. The effect of distinct periodic stimuli is evident in the dissimilar chaotic patterns displayed in the responses. It can be derived from these results that chaotic responses with high similarity originate from very similar periodic stimuli. This agrees in principle that initial representation of a neural response is unique to the stimulus [9, 33].

In this chapter, Hodgkin–Huxley (HH) neural responses generated by varying the inter-spike-interval (ISI) of periodic stimuli are compared to estimate similarity. It is observed that estimating similarity of neural responses exhibiting chaotic dynamics requires knowledge of both firing times and amplitude

distribution. A comparison of similarities estimated by (a) an approach considering firing times alone and (b) an approach based on firing times and amplitude distribution shows that fluctuations induced by periodic stimuli are differentiated better by considering amplitude distribution in addition to firing time information. This paper quantifies the similarity thus estimated and it is observed that it is approximately equal to the percentage number of absolute coincidences. Absolute coincidences are the number of spikes that coincide with respect to both firing times and amplitudes with a pre-defined precision δ.

54.2 Neuronal Model and Synapse

54.2.1 The Neuron Model

The computational model and stimulus for an HH neuron is replicated from [15]. The differential equations of the model are the result of non-linear interactions between the membrane voltage V and the gating variables m, h and n for Na^+ and $K^+$and Cl^-.

$$\left.\begin{aligned} C\frac{dv}{dt} = &- g_{Na}m^3h(V - V_{Na}) - g_K n^4(V - V_K) \\ &- g_L(V - V_L) + I_i \end{aligned}\right\} \tag{54.1}$$

$$\left.\begin{aligned} \frac{dm}{dt} &= -(\alpha_m + \beta_m)m + \alpha_m \\ \frac{dh}{dt} &= -(\alpha_h + \beta_h)h + \alpha_h \\ \frac{dn}{dt} &= -(\alpha_n + \beta_n)n + \alpha_n \end{aligned}\right\} \tag{54.2}$$

$$\left.\begin{aligned} \alpha_m &= 0.1(V + 40)/[1 - e^{-(V+40)/10}] \\ \alpha_h &= 0.07e^{-(V+65)/20} \\ \alpha_n &= 0.01(V + 55)/[1 - e^{-(V+55)/10}] \\ \beta_m &= 4e^{-(V+65)/18} \\ \beta_h &= 1/[1 + e^{-(V+35)/10}] \\ \beta_n &= 0.125e^{-(V+65)/80} \end{aligned}\right\} \tag{54.3}$$

The variable V is the resting potential of the membrane and V_{Na}, V_K and V_L are the reversal potentials of the Na^+, K^+ channels and leakage. The values of the reversal potentials $V_{Na} = 50\,\text{mV}$, $V_K = -77\,\text{mV}$, $V_L = -54.5\,\text{mV}$. The conductance for the ionic channels are $g_{Na} = 120\,\text{mS/cm}^2$, $g_K = 36\,\text{mS/cm}^2$ and $g_L = 0.3\,\text{mS/cm}^2$. The capacitance of the membrane is $C = 1\,\mu\text{F/cm}^2$.

54.2.2 The Synaptic Current

An input spike train given by [34] is considered to generate the pulse component of the external current.

$$U_i(t) = V_a \sum_n \delta(t - t_f) \tag{54.4}$$

where t_f is the firing time and is defined as

$$t_{f(n+1)} = t_{f(n)} + T \tag{54.5}$$

$$t_{f(1)} = 0 \tag{54.6}$$

T represents the ISI of the input spike train and can be varied to generate a different pulse component. The spike train is injected through a synapse to give the pulse current I_P.

$$I_P = g_{syn} \sum_n \alpha(t - t_f)(V_a - V_{syn}) \tag{54.7}$$

g_{syn}, V_{syn} are the conductance and reversal potential of the synapse. The α-function is defined in [32] as

$$\alpha(t) = (t/\tau)e^{-t/\tau}\Theta(t) \tag{54.8}$$

where, τ is the time constant of the synapse and $\Theta(t)$ is the Heaviside step function. $V = 30$ mV, $\tau = 2$ ms, $g_{syn} = 0.5$ mS/cm^2 and $V_{syn} = -50$ mV.

54.2.3 The Total External Current

The total external current applied to the neuron is a combination of static and pulse component

$$I_i = I_S + I_P + \varepsilon \tag{54.9}$$

where I_S is the static and I_P is the pulse current, ε is the random Gaussian noise with zero mean and standard deviation $\sigma = 0.025$.

It is understood that distinct sinusoidal stimuli induce different chaotic oscillations which result in dissimilar neural responses [28–30].

54.3 Similarity Estimation Using $\Gamma_{chaotic}$

The similarity between neural responses exhibiting chaotic oscillations can be determined using $\Gamma_{chaotic}$. $\Gamma_{chaotic}$ estimates similarity through differences between the actual coincidences N_{pcoinc} and the expected number of coincidences $\overline{N_{pcoinc}}$

relative to the average number of spikes in the two spike trains. The similarity is normalised between 0 and 1 by a normalising factor $N_{chaotic}$.

$$\Gamma_{chaotic} = \frac{N_{pcoinc} - \overline{N_{pcoinc}}}{\frac{1}{2}(N_1 + N_2)} \frac{1}{N_{chaotic}} \tag{54.10}$$

where N_{pcoinc} is the number of conditional coincidences (amplitude coincidence given firing time coincidence) between the two spike trains, $\overline{N_{pcoinc}}$ is the conditional mean (average number of amplitude coincidences given firing time coincidences) and $N_{chaotic}$ is the normalising factor for chaotic oscillations. N_1 is the number of spikes in the train 1, N_2 is the number of spikes in train 2. This formulation is based on [32] where similarity based on firing times was estimated through relative number of coincidences without coincidences by chance.

Let $\aleph_1$ and $\aleph_2$ be the normal distributions for the spike trains sp_1 and sp_2 with means μ_1 and μ_2 and respective standard deviations σ_1 and σ_2. The mean probability of coincidence of any amplitude from $\aleph_2$ with an amplitude from $\aleph_1$ can be approximated using the mean of $\aleph_2$.

$$z_{mean} = \frac{\mu_2 - \mu_1}{\sigma_1} \tag{54.11}$$

$$\overline{z_{mean}} = p(z_{mean}) = [z_{mean}]_{from\ Z-table} - [p(\mu_1)]_{from\ Z-table} \tag{54.12}$$

(54.11) and (54.12) give the mean probability that an amplitude from $\aleph_2$ will lie within $\aleph_1$ and coincide with an amplitude from $\aleph_1$. The expected number of amplitude coincidences for any two neural responses generated by periodic stimuli is therefore $\overline{z_{mean}}N_1$. If the rate of fire of sp_2 is ν and the precision for coincidence is δ, then the expected number of coincidences are given by

$$\overline{N_{pcoinc}} = 2\nu\delta N_2 \overline{z_{mean}} \tag{54.13}$$

and the normalising factor $N_{chaotic}$ normalises the estimate of similarity to a value between 0 (dissimilarity) and 1 (exact match)

$$N_{chaotic} = 1 - 2\nu\delta\overline{z_{mean}} \tag{54.14}$$

Similarity on the basis of firing times alone can be determined from (54.10) by omitting the amplitude considerations in N_{pcoinc}, $\overline{N_{pcoinc}}$ and $N_{chaotic}$. This result is consistent with [35].

54.4 Results

Due to the nature of periodic stimuli and chaotic oscillations, estimating similarity between neural responses on the basis of firing times is inaccurate in view of (a) false positives and (b) incorrect inference about stimuli similarity. Similarity

estimation is done for neural responses by varying the stimulus ISI (T) within a limit of 2 ms. Stimulus is varied between 14–16 ms (set I), 13–15 ms (set II) and 15–17 ms (set III) and similarity is estimated by comparing neural responses with reference responses R_{ref} for each set. The reference responses are generated by fixing the ISI (T_{ref}) for the sets at 15 ms for set I, 14 ms for set II and 16 ms for set III. This section compares the similarity estimated by coincidence factor (Γ) and $\Gamma_{chaotic}$.

54.4.1 Comparison of Γ and $\Gamma_{chaotic}$, Set I, 14–16 ms

β represents the difference between the stimulus ISI (T) and a reference ISI T_{ref}. A positive or negative change in β indicates that neural stimuli have dissimilar ISI and their respective influence on neural dynamics is unique to the applied stimulus. The neural responses with underlying chaotic oscillations require both amplitude fluctuations and irregular firing times considered to estimate similarity. It is observed that false positive (circled) obtained by coincidence factor at $\beta = +1$ is eliminated (Fig. 54.1). The overall similarity between pairs of neural responses is reduced in comparison with Γ due to amplitude fluctuations being considered in addition to firing time information.

Table 54.1 gives a clear comparison of firing time and amplitude coincidences. For $\beta = -1$, half of the neural spikes from R_{14} and R_{ref15} coincide with a precision of 2 ms. However, only 20.83% of the amplitudes coincide with a precision 2 mV. This is characteristic of neural responses exhibiting chaotic oscillations—a change in the stimulus reflects on the neural dynamics. Absolute coincidences are conditional coincidences i.e. number of amplitude coincidences given that corresponding firing times coincide. The number of absolute coincidences is 16.67%, which implies that out of all the neural response pairs, only 16.67% exhibit amplitude and firing time coincidences. The similarity estimated by $\Gamma_{chaotic}$ is 0.161. It appears to accurately reflect the absolute coincidences. In addition, $\Gamma_{chaotic}$ also considers coincidences by chance or expected coincidences which renders the similarity estimated by $\Gamma_{chaotic}$ unique to a pair of neural responses.

At $\beta = 0$, both stimuli have the same ISI ($T = 15\,\text{ms}, T_{ref} = 15\,\text{ms}$). All neural spikes coincide in firing times and amplitudes. The absolute coincidences confirm that the neural responses are an exact match, hence, similarity $\Gamma_{chaotic} = 1$. At $\beta = +1$, as all neural spikes show firing time coincidences, coincidence factor classifies the neural responses R_{16} and R_{ref15} as identical. This result is a false positive as indicated by the number of amplitude fluctuations. Though all neural spikes coincide with firing times, only 29.17% of amplitudes coincide, hence the absolute coincidences are 29.17%. The corresponding value of $\Gamma_{chaotic}$ is 0.2729 which is substantially lower than 1 (estimated by coincidence factor). The consideration of amplitude fluctuations in addition to firing time information successfully eliminates the false positive.

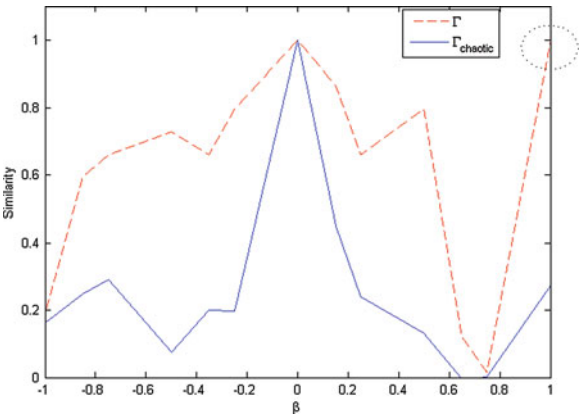


Fig. 54.1 Similarity of neural responses generated by periodic stimuli with $14\,\text{ms} \leq T \leq 16\,\text{ms}$ and $T_{ref} = 15\,\text{ms}$. Γ represents the similarity estimated by coincidence factor and $\Gamma_{chaotic}$ is the similarity based on firing times and amplitudes coincidences. The incorporation of amplitude fluctuations to estimate similarity helps $\Gamma_{chaotic}$ eliminate false positive (*circled*) at $\beta = +1$

Table 54.1 Firing time, amplitude and absolute coincidences for various values of β in set I

β	Firing time coincidences (%)	Amplitude coincidences (%)	Absolute coincidences (%)	$\Gamma_{chaotic}$
−1	50	20.83333	16.6667	0.161
	75	29.1677	25	0.2489
	79.1667	37.5	29.1667	0.2908
	83.3333	12.5	8.3333	0.0753
	79.1667	25	20.8333	0.1997
	87.5	29.1667	20.8333	0.1984
	100	100	100	1
	91.6667	50	45.8333	0.4513
	79.1667	33.3333	25	0.2392
	87.5	16.6667	16.6667	0.1312
	37.5	12.5	4.1667	0.0003
+1	100	29.1667	29.1667	0.2729

$\Gamma_{chaotic}$ represents the similarity between pairs of neural responses. Firing time coincidence precision is 2 ms, amplitude coincidence precision is 2 mV and absolute coincidence is a conditional coincidence of amplitudes given that corresponding firing times coincide. $\Gamma_{chaotic}$ accurately calculates similarity and this can be correlated with the percentage of absolute coincidences

54.4.2 Comparison of Γ and $\Gamma_{chaotic}$, Set II, 13–15 ms

False positives (circled) estimated by coincidence factor, Γ, shown in Fig. 54.2 occur for $-0.5 \leq \beta < 0$. Table 54.2 shows that firing time coincidences are 100% for $-0.5 \leq \beta < 0$. These pairs of neural responses are classified identical by coincidence factor. However, the corresponding amplitude coincidences are 12.5, 33.33 and 29.17% which indicate that though the firing times coincide, the amplitude fluctuations are not identical, hence these are termed as false positives.

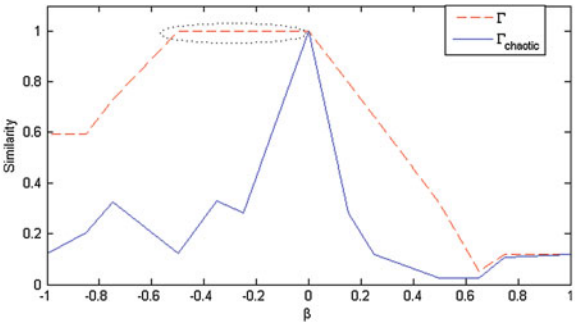


Fig. 54.2 Similarity of neural responses generated by periodic stimuli with $13\,\text{ms} \leq T \leq 15\,\text{ms}$ and $T_{ref} = 14\,\text{ms}$. Γ represents the similarity estimated by coincidence factor and $\Gamma_{chaotic}$ is the similarity based on firing times and amplitudes coincidences. $\Gamma_{chaotic}$ eliminates false positives (*circled*) between $-0.5 \leq \beta < 0$

Table 54.2 Firing time, amplitude and absolute coincidences for various values of β in set II

β	Firing time coincidences (%)	Amplitude coincidences (%)	Absolute coincidences (%)	$\Gamma_{chaotic}$
−1	75	16.6667	12.5	0.1204
	75	25	20.8333	0.2044
	83.3333	37.5	33.3333	0.3272
	100	12.5	12.5	0.1224
	100	33.3333	33.3333	0.3302
	100	29.1667	29.1667	0.2846
	100	100	100	1
	87.5	37.5	29.1667	0.2857
	79.1667	16.6667	12.5	0.1178
	58.3333	8.3333	4.1667	0.0263
	41.6667	33.3333	4.1667	0.0243
	45.8333	45.8333	12.5	0.1075
+1	45.8333	20.8333	12.5	0.1188

$\Gamma_{chaotic}$ eliminates the false positives occurring for $-0.5 \leq \beta < 0$. Similarity estimated by $\Gamma_{chaotic}$ correlates to the percentage of absolute coincidences

These non-identical amplitude fluctuations are caused by the dissimilar periodic stimuli. The corresponding absolute coincidences for the false positives are 12.5, 33.33 and 29.17%. The similarity estimated by $\Gamma_{chaotic}$ is 0.1224, 0.3302 and 0.2846 respectively which reflects the absolute coincidences. The composite consideration of irregular firing times and varying amplitudes helps differentiate neural responses and relate their dissimilarity to the stimuli.

For $\beta = 0$, the neural responses are generated by identical stimuli ($T = 14\,\text{ms}$ and $T_{ref} = 14\,\text{ms}$). Identical stimuli cause similar chaotic oscillations, hence the resulting neural responses are an exact match. This is seen in Table 54.2 and Fig. 54.2, at $\beta = 0$, the firing time coincidences, amplitude coincidences and the absolute coincidences are 100%. This justifies that the neural responses are

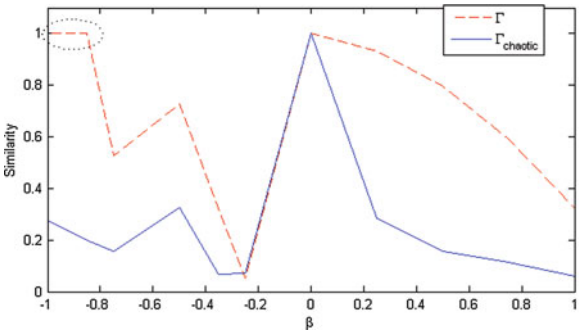


Fig. 54.3 Similarity of neural responses generated by periodic stimuli with $15\,\text{ms} \leq T \leq 17\,\text{ms}$ and $T_{ref} = 16\,\text{ms}$. Γ represents the similarity estimated by coincidence factor and $\Gamma_{chaotic}$ is the similarity based on firing times and amplitudes. The incorporation of amplitude fluctuations to estimate similarity helps $\Gamma_{chaotic}$ eliminate false positives (*circled*) for $-1 \leq \beta \leq -0.75$

Table 54.3 Firing time, amplitude and absolute coincidences for various values of β in set III

β	Firing time coincidences (%)	Amplitude coincidences (%)	Absolute coincidences (%)	$\Gamma_{chaotic}$
−1	100	29.1667	29.1667	0.2754
	100	20.8333	20.8333	0.1992
	70.8333	25	16.6667	0.1581
	83.3333	33.3333	33.3333	0.3243
	58.3333	12.5	8.3333	0.0666
	41.6667	12.5	8.3333	0.0704
	100	100	100	1
	95.8333	33.3333	29.1667	0.2851
	87.5	20.8333	16.6667	0.1558
	75	16.6667	12.5	0.1124
+1	58.3333	37.5	8.3333	0.0617

The false positives determined by coincidence factor for $-1 \leq \beta \leq -0.75$ are eliminated. Similarity between neural response pairs estimated by $\Gamma_{chaotic}$ correlates with the percentage of absolute coincidences

an exact match and they were generated by identical stimuli, hence $\Gamma_{chaotic} = 1$. The similarity determined by $\Gamma_{chaotic}$ for other neural response pairs is also consistent in correlation with the absolute coincidences.

54.4.3 Comparison of Γ and $\Gamma_{chaotic}$, Set III, 15–17 ms

Set III exhibits false positives (circled) for $-1 \leq \beta \leq -0.75$ (Fig. 54.3). Table 54.3 shows that the corresponding firing time coincidences are 100% which result in coincidence factor classifying the pair of neural responses identical. However, the amplitude coincidences are 29.17 and 20.83% indicating that the

underlying oscillations are non-identical. The corresponding similarity determined by $\Gamma_{chaotic}$ is 0.2754 and 0.1992. As $\Gamma_{chaotic} \neq 1$, the neural responses are not an exact match and they were generated by dissimilar stimuli.

For $\beta = 0$, the neural responses are generated by identical stimuli ($T = 16$ ms and $T_{ref} = 16$ ms). The similarity between other neural response pairs reflects their dissimilar stimuli and correlates with the absolute coincidences. For $\beta > 0$, $\Gamma_{chaotic}$ decreases with an increase in β indicating that similarity between the neural responses decreases with an increase in the difference in the ISI of two stimuli. Any difference in the ISI of a stimulus causes a temporal change and effect on neural dynamics is evident in the dissimilarity estimated by $\Gamma_{chaotic}$.

54.5 Conclusions

The nature of a periodic stimulus is responsible to induce chaotic oscillations in a biological neuron. Irregular inter-spike interval (ISI) and fluctuating amplitudes are the characteristics of chaotic oscillations absent in steady state responses generated by constant current stimuli. Information on stimuli similarity can be derived from neural response comparison. Estimating similarity based on firing times (coincidence factor) alone is insufficient in view of (a) false positives and (b) incorrect inference about neural stimuli. Firing time information is adequate to estimate similarity of steady state responses; however, in the presence of chaotic oscillations or when the amplitudes of a neural response fluctuate, amplitude and firing time collectively reflect the true dynamics of a neuron and therefore both should feature in similarity estimation [28–30].

The amplitudes a neural response exhibiting chaotic oscillations fit a Normal distribution and using the properties of Normal distribution, it is possible to determine amplitude coincidences. Similarity between these responses can be estimated by a composite similarity measure based on amplitude and firing time coincidences. In addition, similar periodic stimuli induce similar chaotic patterns in the neural responses and therefore the resulting neural responses have a high degree of similarity. The effect of distinct periodic stimuli is evident in the dissimilar chaotic patterns displayed in the responses. It follows that chaotic responses with high similarity originate from very similar periodic stimuli. This agrees in principle that initial representation of a neural response is unique to the stimulus [9, 33].

The results show that the similarity estimated using both firing times and amplitudes can be quantified by analyzing the number of absolute coincidences. Absolute coincidences are the number of spikes that coincide with respect to both firing times and amplitudes with a pre-defined precision δ. It is observed that similarity estimated by $\Gamma_{chaotic}$ is approximately equal to the percentage number of absolute coincidences. If the number of absolute coincidences are 25%, then the similarity estimated by $\Gamma_{chaotic}$ is approximately 0.25. This quantification ensures that the estimated similarity is realistic and mathematically realizable.

References

1. Lundström I (1974) Mechanical wave propagation on nerve axons. J Theoret Biol 45: 487–499
2. Abbott LF, Kepler TB (1990) Model neurons: from Hodgkin Huxley to Hopfield. In: Garrido L (ed) Statistical mechanics of neural networks. Springer, Berlin, pp 5–18
3. Hasegawa H (2000) Responses of a Hodgkin–Huxley neuron to various types of spike-train inputs. Phys Rev E 61(1):718–726
4. Agüera y Arcas B, Fairhall AL (2003) What causes a neuron to spike? Neural comput 15:1789–1807
5. Agüera y Arcas B, Fairhall AL, Bialek W (2003) Computation in a single neuron: Hodgkin and Huxley revisited. Neural Comput 15:1715–1749
6. Fourcaud-Trocmé N, Hansel D, van Vreeswijk C, Brunel N (2003) How spike generation mechanisms determine the neuronal response to fluctuating inputs. J Neurosci 23(37): 11628–11640
7. Kepecs A, Lisman J (2003) Information encoding and computation with spikes and bursts. Network Comput Neural Syst 14:103–118
8. Bokil HS, Pesaran B, Andersen RA, Mitra PP (2006) A method for detection and classification of events in neural activity. IEEE Trans Biomedical Eng 53(8):1678–1687
9. Davies RM, Gerstein GL, Baker SN (2006) Measurement of time-dependent changes in the irregularity of neural spiking. J Neurophysiol 96:906–918
10. Diba K, Koch C, Segev I (2006) Spike propagation in dendrites with stochastic ion channels. J Comput Neurosci 20:77–84
11. Dimitrov AG, Gedeon T (2006) Effects of stimulus transformations on estimates of sensory neuron selectivity. J Comput Neurosci 20:265–283
12. Izhikevich EM (2006) Polychronization: computation with spikes. Neural Comput 18: 245–282
13. Li X, Ascoli GA (2006) Comput simulation of the input–output relationship in hippocampal pyramidal cells. J Comput Neurosci 21:191–209
14. Rinzel J (1985) Excitation dynamics: insights from simplified membrane models. Theoret Trends Neurosci Federal Proc 44(15):2944–2946
15. Panzeri S, Schultz SR, Treves A, Rolls ET (1999) Correlations and the encoding of information in the nervous system. Proc R Soc Lond B 266:1001–1012
16. Gabbiani F, Metzner W (1999) Encoding and processing of sensory information in neuronal spike trains. J Biol 202:1267–1279
17. Wang XJ, Buzsáki G (1996) Gamma oscillation by synaptic inhibition in a hippocampal interneuronal network model. J Neurosci 16(2):6402–6413
18. Guttman R, Feldman L, Jakobsson E (1980) Frequency entrainment of squid axon membrane. J Membrane Biol 56:9–18
19. Matsumoto G, Kim K, Ueda T, Shimada J (1980) Electrical and computer simulations upon the nervous activities of squid giant axons at and around the state of spontaneous repetitive firing of action potentials. J Phys Soc Jpn 49:906
20. Aihara K, Matsumoto G, Ikegaya Y (1984) Periodic and non-periodic responses of a periodically forced Hodgkin–Huxley oscillator. J Theoret Biol 109:249–269
21. Matsumoto G, Aihara K, Ichikawa M, Tasaki A (1984) Periodic and nonperiodic responses of membrane potentials in squid giant axons during sinusoidal current simulations. J Theoret Neurobiol 3:1–14
22. Hayashi H, Ishizuka S, Hirakawa K (1985) Chaotic responses of the pacemaker neuron. J Phys Soc Jpn 54:2337
23. Holden AV (1987) Chaos. Manchester University Press, Manchester
24. Hasegawa H (2008) Information conveyed by neuron populations—firing rate, fluctuations and synchrony. Neuroquantology 6(2):105–118
25. Kaplan D, Glass L (1995) Understanding non linear dynamics. Springer, New York

26. Wilson HR (1999) Spikes, decisions and actions—the dynamical foundations of neuroscience. Oxford University Press Inc, New York
27. Hodgkin A, Huxley A (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 117:500–544
28. Sarangdhar M, Kambhampati C (2008) Spiking neurons: is coincidence-factor enough to compare responses with fluctuating membrane voltage? In: World Congress on engineering 2008: the 2008 international conference of systems biology and bioengineering, vol 2, London, UK, 2–4 July 2008, pp 1640–1645
29. Sarangdhar M, Kambhampati C (2008) Spiking neurons and synaptic stimuli: determining the fidelity of coincidence-factor in neural response comparison. Eng Lett 16(4):512–517
30. Sarangdhar M, Kambhampati C (2009) Spiking neurons and synaptic stimuli—neural response comparison using coincidence-factor. In: Gelman L, Balkan N, Ao S (eds) Advances in electrical engineering and computational science. Springer, Berlin
31. Joeken S, Schwegler H (1995) Predicting spike train responses in neuron models. In: Verleysen M (ed) Proceedings of the 3rd European symposium on artificial neural networks 1995, Brussels, Belgium, April 19–21, pp 93–98
32. Kistler WM, Gerstner W, Leo van Hemmen J (1997) Reduction of the Hodgkin–Huxley equations to a single-variable threshold model. Neural Comput 9:1015–1045
33. Chechik G, Anderson MJ, Bar-Yosef O, Young ED, Tishby N, Nelken I (2006) Reduction of information redundancy in the ascending auditory pathway. Neuron 51:359–368
34. Park MH, Kim S (1996) Analysis of phase models for two coupled Hodgkin–Huxley neurons. J Korean Phys Soc 29(1):9–16
35. Jolivet R, Lewis TJ, Gerstner W (2004) Generalized integrate-and-fire models of neuronal activity approximate spike trains of a detailed model to a high degree of accuracy. J Neurophysiol 92:959–976

Chapter 55
Reduction of HIV Infection that Includes a Delay with Cure Rate During Long Term Treatment: A Mathematical Study

Priti Kumar Roy and Amar Nath Chatterjee

Abstract Progress towards antiviral treatment of HIV infected individual has largely been improved in recent years. From amassed literature it has been observed that antiviral treatment during disease progression can develop CTL and these CTL has an immense importance to control the disease progression. During long term treatment, CD4 and CD8 mediated immune response take part in an effective role against HIV and thus virus load is abridged monotonically or else high virus load leads to have weak immunity. Here we extended our work (Roy PK, Chatterjee AN, Chattopadhyay B (2010) Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010:533–538) and further introduced a population model representing long term dynamics of HIV infection in response to available drug therapies considering a cure rate and a discrete time delay in the disease transmission term. We studied the model in different avenue. Our studies reveal that delay in the disease transmission term competes with the killing rate of infected T-cells as well as stimulation rate of CTL. Increasing of delay in this case makes the system progressively unstable and when delay reach to its threshold value, a periodical solution arise through Hopf bifurcation. It is also been observed that increasing value of cure rate could be able to improvise towards stability of the system inspite of delay effect.

P. K. Roy (✉) · A. N. Chatterjee
Centre for Mathematical Biology and Ecology, Department of Mathematics, Jadavpur University, Kolkata 700032, India
e-mail: pritiju@gmail.com

A. N. Chatterjee
e-mail: anchaterji@gmail.com

S. I. Ao and L. Gelman (eds.), *Electrical Engineering and Applied Computing*, Lecture Notes in Electrical Engineering, 90, DOI: 10.1007/978-94-007-1192-1_55,
© Springer Science+Business Media B.V. 2011

55.1 Introduction

Several studies have been articulated that HIV is a retrovirus and it target lymphocyte cell mainly CD4$^+$T cells. The human immunodeficiency viruses (HIV-1 and HIV-2) impair our immune system. When CD4$^+$T cell count reduces from $1,000$ to $200\,\text{mm}^{-3}$ or below, the cell medicated immune system is lost and the body becomes progressively infectious [1–7].

In recent years antiviral treatment has largely improved. It has been observed that during long term treatment of drugs (combination of RTI and PI) reduces the virus load 10–100 fold and 25% increases in CD4$^+$T*T* cell count [6]. It has also been observed that after commence of the treatment virus attains a new steady state approximately 10–100 fold lower than the base line value before the treatment [6].

Here we consider a mathematical model of HIV during long term treatment. We are interested to see the drug induced changes in long term drug therapy. In this model free virus population is omitted as because at steady state, free virus population is proportional to the virus producing cells. During long term treatment, system generates CTL and this CTL acts against virus producing cells and kill them.

HIV is thought to be primarily a non cytopathic virus, primarily the virus producing cell are lost either through death, mainly immune mediated killing or via cure that is loss of cccDNA [8]. Secondly due to antiviral therapy some infected T-cells are killed by CTL [6]. Here we suggest both cytolytic and non cytolytic mechanisms of infected cell so that model become more realistic and accurate.

Here we also consider a time lag or delay in the disease transmission process. Since the process of infection between uninfected and infected CD4$^+$T cells is not an instantaneous we formulate a delay differential equation considering a delay ($\tau > 0$) in the disease transmission term in our system. In this paper we solve the model analytically and numerically. In our analytical study we have found that there are two steady state—uninfected and infected steady state. We have also shown that these stability of two steady state depends on the basic reproduction ratio.

We also examine that the effect of delay for the model and find out the condition for stability. Here we see that for certain parametric restriction delay model remain asymptotically stable. Numerical simulation confirmed our analytical results.

55.1.1 Presentation of ODE Model

To formulate a mathematical model, we first consider the simplest T-cell dynamics [6] given as

$$\begin{aligned} \dot{x} &= \lambda - d_1 x - \beta_1 xy, \\ \dot{y} &= \beta_1 xy - d_2 y, \end{aligned} \tag{55.1}$$

where λ is the constant rate of production (or supply) of immune competent T-cell from the thymus, d_1 is the natural death rate of infectible cell, d_2 is the death rate of virus producing cell, and β_1 is the rate of infection of uninfected cells.

Here we are interested to see the drug induced changes at the steady state. Thus the free virus population is not considered as it is proportional to the virus producing cell population [6].

During short term dynamics the effect of immune response is negligible or remain constant. But when HAART is introduced viral load reduces and the effect of immune response is activated. Then the extended model dynamics becomes [6].

$$\begin{aligned} \dot{x} &= \lambda - d_1 x - \beta_1 xy, \\ \dot{y} &= \beta_1 xy - d_2 y - \beta_2 yz, \\ \dot{z} &= sy - d_3 z, \end{aligned} \tag{55.2}$$

where β_2 is the killing rate of virus producing cell by CTL, s is the rate of stimulation of CTL and d_3 is the death rate of CTL. Since HIV is thought to be primarily a non cytopathtic virus, thus the virus producing cells are lost either immune mediated killing or loss of cccDNA (i.e. cure). Thus we reconstruct the model (55.2) of HIV infection of $CD4^+T$ cells with cure rate. Then the model becomes

$$\begin{aligned} \dot{x} &= \lambda - d_1 x - \beta_1 xy + \delta y, \\ \dot{y} &= \beta_1 xy - (d_2 + \delta) y - \beta_2 yz, \\ \dot{z} &= sy - d_3 z \end{aligned} \tag{55.3}$$

with the initial condition:

$$x(0) > 0, \quad y(0) > 0, \quad z(0) > 0. \tag{55.4}$$

Here δ is the cure rate that is the non cytotoxic loss of virus producing cells. Thus the total disappearance of infected cells is $(d_2 + \delta)$.

55.2 Equilibria and Local Stability

The system (55.3) together with (55.4) possess the following positive equilibrium:

(i) an uninfected steady state $E_0 = (x_0, 0, 0)$

(ii) an infected steady state $\bar{E} = (\bar{x}, \bar{y}, \bar{z})$.

where,

$$x_0=\frac{\lambda}{d_1},$$
$$\bar{x}=\frac{(2\beta_1\delta d_3-\beta_2 d_1 s+\beta_1 d_2 d_3)+\sqrt{(2\beta_1\delta d_3-\beta_2 d_1 s+\beta_1 d_2 d_3)^2-4{\beta_1}^2 d_3(\delta^2 d_3+d_2 d_3\delta-\lambda\beta_2 s)}}{2{\beta_1}^2 d_3},$$
$$\bar{y}=\frac{d_3(\beta_1\bar{x}-d_2-\delta)}{s\beta_2},\ \bar{z}=\frac{\beta_1\bar{x}-d_2-\delta}{\beta_2} \tag{55.5}$$

satisfying the following inequality $\delta<\frac{2\lambda\beta_1-d_1 d_2}{2d_1}$. Thus we get a critical value of cure rate i.e. $\delta_{crit}=\frac{2\lambda\beta_1-d_1 d_2}{2d_1}$.

Let R_0 be the basic reproduction ratio of the system, which represents the average number of secondary infected caused by a single infected cell in an entirely susceptible cell population through out its infection period. For the system (55.3) $R_0=\frac{\lambda\beta_1 s}{d_1 d_3(d_2+\delta)}$.

Theorem 55.1 *If* $R_0<1$, $E_0=(x_0,0,0)$ *is locally stable, and if* $R_0>1$, $E_0=(x_0,0,0)$ *is unstable.*

Theorem 55.2 *If* (i) $R_0>1$, (ii) $\delta<\delta_{crit}$, *and* (iii) $A_1A_2-A_3>0$ *then the infected steady state* $\bar{E}=(\bar{x},\bar{y},\bar{z})$ *is locally asymptotically stable.*

Proof To comment upon the existence of the local stability of the infected steady state $\bar{E}$ for $\delta<\delta_{crit}$, we consider the linearized system of (55.3) at $\bar{E}$. The Jacobian matrix at $\bar{E}$ is given by

$$\begin{pmatrix} -d_1-\beta_1\bar{y} & -\beta_1\bar{x}+\delta & 0 \\ \beta_1\bar{y} & \beta_1\bar{x}-(d_2+\delta+\beta_2\bar{z}) & -\beta_2\bar{y} \\ 0 & s & -d_3 \end{pmatrix}. \tag{55.6}$$

Then the characteristic equation of the system (55.6) becomes

$$\rho^3+A_1\rho^2+A_2\rho+A_3=0, \tag{55.7}$$

where,

$$\begin{aligned} A_1&=d_1+d_2+d_3+\delta+\beta_1(\bar{y}-\bar{x})+\beta_2\bar{z},\\ A_2&=d_3(d_1+d_2+\delta+\beta_2\bar{z}-\beta_1\bar{x})+\beta_2 s\bar{y}+d_1(d_2+\delta)\\ &\quad+\beta_1(d_1\bar{z}-d_1\bar{x}+d_2\bar{y}+\beta_1\bar{y}\bar{z}),\\ A_3&=d_1 d_3(d_2+\delta+\beta_2\bar{z}-\beta_1\bar{x})\\ &\quad+\bar{y}(\beta_2 s d_1+\beta_1 d_2 d_3+\beta_1\beta_2 d_3\bar{z}+s\beta_1\beta_2\bar{y}). \end{aligned} \tag{55.8}$$

By Routh–Hurwitz criterion, the necessary and sufficient condition for locally asymptotic stability of the steady state are $A_1>0$, $A_3>0$, $A_1A_2-A_3>0$. □

55.3 Boundedness and Permanence of the System

In this section we first analyze that the system is bounded. Here we assume a positively invariant set $\Gamma = \{(x(t), y(t), z(t) | x(t) > 0, y(t) > 0, z(t) > 0\}$. Also we assume that $x(t), y(t)$ and $z(t)$ are random positive solution of the system with initial values.

Theorem 55.3 *The system* (55.3) *is bounded above for large value of* $M > 0$ *i.e.* $x(t) \leq M,\ y(t) \leq M,\ z(t) \leq M$ *for large* $t \geq T$.

Proof Let $U(t) = x(t) + y(t)$.
Now, $\dot{U} = \dot{x} + \dot{y} = \lambda - (d_1 x + d_2 y) - \beta_2 yz \leq \lambda - d_1(x + y)$.

So the uninfected and infected cell population are always bounded. Thus there exist $M > 0$ such that $x(t) \leq M,\ y(t) \leq M,\ z(t) \leq M$ for large $t \geq T$. From the third equation of (55.3) it is easy to see that $z(t)$ is also bounded. Hence it is proved that system is bounded one. □

Theorem 55.4 *The system* (55.3) *is bounded below for any lower value of* $m > 0$ *i.e.* $x(t) \geq m,\ y(t) \geq m,\ z(t) \geq m$ *for large* $t \geq T$.

Proof To prove the theorem we choose large $t \geq T$ such that $\dot{y} = y(\beta_1 x - d_1 - \delta - \beta_2 z) \geq y(\beta_1 x - d_1 - \delta - \beta_2 M) \geq 0$ and

$\dot{z} = sy - d_3 z \geq sy - d_3 M \geq 0$

for $x \geq m_1$ and $y \geq m_2$

where $m_1 = \frac{d_1 + \delta + \beta_2 M}{\beta_1}$ and $m_2 = \frac{d_3 M}{s}$. Then $z(t)$ is also bounded below, i.e. $z(t) \geq m_3$ where $m_3 = \frac{\lambda \beta_1 - d_1(d_1 + \delta + \beta_2 M)}{\beta_1 d_3 (d_1 + \beta_2 M)}$.

Then there exist $m = \max\{m_1, m_2, m_3\}$, such that $x(t) \geq m,\ y(t) \geq m,\ z(t) \geq m$ for large $t \geq T$.

Hence it is proved that the system is bounded below. Thus we can define a positive invariant set $\Gamma = \{(x(t), y(t), z(t)) | m \leq x(t) \leq M, m \leq y(t) \leq M, m \leq z(t) \leq M\}$, where each solution of the system (55.3) with positive initial value will be enter in the compact region Γ and remaining finally. Summarizing the above analysis we can establish the theorem stated below. □

Theorem 55.5 *The positive invariant solution* Γ *of the system* (55.3) *with boundedness is permanent.*

55.4 Global Stability of the System

Theorem 55.6 *For the system* (55.3), $\bar{E}$ *is globally asymptotically stable in* Γ *if* $R_0 > 1$ *together with the condition* $A_1 A_2 - A_3 > 0$.

Proof To prove the Global stability of the system we construct the Lyapunov function.

$V(x(t), y(t), z(t)) = \frac{w_1}{2}(x - \bar{x})^2 + w_2(y - \bar{y} - \ln\frac{y}{\bar{y}}) + \frac{w_3}{2}(z - \bar{z})^2$.

Calculating the upper right derivative of $V(x(t), y(t), z(t))$ along the system (55.3) we obtain,

$$D^+V(x(t), y(t), z(t)) = w_1(x - \bar{x})\dot{x} + w_2\frac{(y - \bar{y})}{y}\dot{y} + w_3(z - \bar{z})\dot{z}$$
$$= -[w_1(d_1 + \beta_1 y)(x - \bar{x})^2 + w_3 d_3(z - \bar{z})^2]$$
$$+ \{w_1(\beta_1\bar{x} - \delta) - w_2\beta_1\}(x - \bar{x})(y - \bar{y})$$
$$+ (w_3 s - w_2\beta_2)(y - \bar{y})(z - \bar{z}).$$

If $w_1(\beta_1\bar{x} - \delta) = w_2\beta_1$ and $w_2\beta_2 = w_3 s$, then we have

$$D^+V(x(t), y(t), z(t)) = -[w_1(d_1 + \beta_1 y)(x - \bar{x})^2 + w_3 d_3(z - \bar{z})^2] < 0.$$

Hence we can say that the system $\bar{E}$ is global asymptotically stable if $R_0 > 1$. □

55.5 Delay Model

In this section we introduce a time delay into Eq. 55.3 on the assumption that the transmission of the disease is not an instantaneous process. In reality there is a time lag between the process of infection of cells to cells become actively infected. Thus the model becomes

$$\begin{aligned}
\dot{x} &= \lambda - d_1 x - \beta_1 x(t - \tau)y(t - \tau) + \delta y,\\
\dot{y} &= \beta_1 x(t - \tau)y(t - \tau) - (d_2 + \delta)y - \beta_2 yz,\\
\dot{z} &= sy - d_3 z,
\end{aligned} \tag{55.9}$$

with the initial condition

$$x(\theta) = x_\theta, \quad y(\theta) = y_\theta, \quad z(\theta) = z_\theta, \quad \theta \in [-\tau, 0]. \tag{55.10}$$

55.6 Analysis

In this section we also study the local and global stability of the disease free equilibrium E_0 of the delay differential equation. Here we consider the two stability function when $R_0 < 1$ and $R_0 > 1$. For delayed system we also find two steady state $E_0 = (x_0, 0, 0)$ and $\bar{E} = (\bar{x}, \bar{y}, \bar{z})$ similar to the non delayed system. In previous section we have seen that for $R_0 > 1, E_0$ is unstable. The result is same as before for delayed system. Then E_0 is unstable if $R_0 > 1$ and the system moves towards infected steady state $\bar{E}$. To study the stability of the steady state $\bar{E}$,

we linearized the system by substituting $X(t) = x(t) - \bar{x},\ Y(t) = y(t) - \bar{y},\ Z(t) = z(t) - \bar{z}$. Then the linearized system of Eq. 55.9 at $\bar{E}$ is given by,

$$\begin{aligned}
\dot{X} &= -d_1 X - \beta_1 \bar{y} X(t-\tau) - \beta_1 \bar{x} Y(t-\tau) + \delta Y, \\
\dot{Y} &= \beta_1 \bar{y} X(t-\tau) + \beta_1 \bar{x} Y(t-\tau) - (d_2 + \delta) Y - \beta_2 \bar{z} Y - \beta_2 \bar{y} Z, \\
\dot{Z} &= sY - d_3 Z.
\end{aligned} \tag{55.11}$$

The linearized system can express in a matrix form as follows $\dot{W(t)} = J_1 W(t) + J_2 W(t-\tau)$. where,

$$J_1 = \begin{pmatrix} -d_1 & \delta & 0 \\ 0 & -(d_2 + \delta + \beta_2 \bar{z}) & -\beta_2 \bar{y} \\ 0 & s & -d_3 \end{pmatrix}$$

and

$$J_2 = \begin{pmatrix} -\beta_1 \bar{y} & -\beta_1 \bar{x} & 0 \\ \beta_1 \bar{y} & \beta_1 \bar{x} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and $W(\cdot) = ((X(\cdot), Y(\cdot), Z(\cdot))^T$.

The characteristic equation of the system (55.11) is given by

$$\rho^3 + a_1 \rho^2 + a_2 \rho + a_3 + (a_4 \rho^2 + a_5 \rho + a_6) e^{-\rho\tau} = 0, \tag{55.12}$$

where,

$$\begin{aligned}
a_1 &= d_1 + d_2 + d_3 + \delta + \beta_2 \bar{z}, \\
a_2 &= (d_1 + d_3)(d_2 + \delta + \beta_2 \bar{z}) + d_3(d_1 + \beta_2 s \bar{y}), \\
a_3 &= d_1 d_3 (d_2 + \delta + \beta_2 \bar{z}) + \beta_2 d_1 s \bar{y}, \\
a_4 &= \beta_1 (\bar{y} - \bar{x}), \\
a_5 &= \beta_1 \{\bar{y}(d_2 + d_3 + \beta_2 \bar{z}) - \bar{x}(d_1 + d_3)\}, \\
a_6 &= \beta_1 \{\bar{y}(d_2 d_3 + \beta_2 d_3 \bar{z} + \beta_2 s \bar{y}) - d_1 d_2 \bar{x}\}.
\end{aligned} \tag{55.13}$$

We know that the infected steady state is asymptotically stable if all roots of the characteristic equation have negative real parts. Since Eq. 55.12 is a transcendental equation and it has infinitely many eigen values, thus it is very difficult to deal with this equation. From classical Routh–Hurwitz criterion we cannot discuss the characteristic equation. By Rouche's theorem and continuity in τ the characteristic equation (55.12) has roots with positive real parts if and only if it has purely imaginary roots, then we can find out the condition for all eigen values to have negative real parts.

Let $\rho = u(\tau) + iv(\tau),\ (v > 0)$ be the eigen value of the characteristic equation (55.12) where u and v are depends on τ. For non delayed system $(\tau = 0)$, $\bar{E}$ is stable if $u(0) < 0$. Since τ is continuous then for small value of $\tau > 0$, we still have

$u(\tau)<0$ and $\bar{E}$ remain stable. If for certain $\tau_0 > 0, u(\tau_0) = 0$ then the steady state $\bar{E}$ losses its stability and it becomes unstable when $u(\tau) > 0$ for $\tau > \tau_0$. Also if all the roots of the characteristic equation (55.12) stand real (i.e. $v(\tau) = 0$), then $\bar{E}$ is always stable. Now to see whether the equation (55.12) has purely imaginary roots or not, we put $\rho = iv$ in Eq. 55.12 and separating the real and imaginary parts. We have,

$$a_1 v^2 - a_3 = (a_6 - a_4 v^2)\cos(v\tau) + a_5 v \sin(v\tau), \tag{55.14}$$

$$v^3 - a_2 v = a_5 v \cos(v\tau) - (a_6 - a_4 v^2)\sin(v\tau). \tag{55.15}$$

Squaring and adding the above equation we get

$$v^6 + ({a_1}^2 - 2a_2 - {a_4}^2)v^4 + ({a_2}^2 - {a_5}^2 + 2a_4 a_6 - 2a_1 a_3)v^2 + ({a_3}^2 - {a_6}^2) = 0. \tag{55.16}$$

Let
$\xi = v^2, \alpha_1 = {a_1}^2 - 2a_2 - {a_4}^2, \alpha_2 = {a_2}^2 - {a_5}^2 + 2a_4 a_6 - 2a_1 a_3, \alpha_3 = {a_3}^2 - {a_6}^2.$
Then the Eq. 55.16 becomes

$$F(\xi) = \xi^3 + \alpha_1 \xi^2 + \alpha_2 \xi + \alpha_3 = 0. \tag{55.17}$$

Since $\alpha_3 = {a_3}^2 - {a_6}^2 > 0$ and $\alpha_2 > 0$ then the Eq. 55.17 has no positive real roots. Now,

$$F'(\xi) = 3\xi^2 + 2\alpha_1 \xi + \alpha_2 = 0. \tag{55.18}$$

Then the roots of the Eq. 55.18 are

$$\xi_1 = \frac{-\alpha_1 + \sqrt{{\alpha_1}^2 - 3\alpha_2}}{3}, \quad \xi_2 = \frac{-\alpha_1 - \sqrt{{\alpha_1}^2 - 3\alpha_2}}{3}. \tag{55.19}$$

Since $\alpha_2 > 0$, then $\sqrt{{\alpha_1}^2 - 3\alpha_2} < \alpha_1$. Hence both the root are negative. Thus the Eq. 55.18 does not have any positive roots. Since $F(0) = \alpha_3 \geq 0$, so the Eq. 55.17 have no positive roots. In brief we can get the *Theorem 55.7*.

Theorem 55.7 *If the system satisfy* (i) $R_0 > 1$ (ii) $A_1 A_2 - A_3 > 0$ *and* (iii) $\alpha_3 \geq 0$ *and* $\alpha_2 > 0$, *then the infected steady state is asymptotically stable for all* $\tau \geq 0$. *Now if the above condition are not satisfied, i.e. if* (i) $\alpha_3 < 0$, *then* $F(0) < 0$ *and* $\lim_{\xi \to \infty} F(\xi) = \infty$. *Thus Eq. 55.18 has at least one positive root say* ξ_0, *then Eq. 55.17 has at least one positive root say* v_0. *If* $\alpha_2 < 0$ *then* $\sqrt{{\alpha_1}^2 - 3\alpha_2} > \alpha_1$, *then from* (55.19), $\xi_1 = \frac{-\alpha_1 + \sqrt{{\alpha_1}^2 - 3\alpha_2}}{3} > 0$. *Hence the positive root* v_0 *exist for* (55.16). *Thus the characteristic equation* (55.16) *has a pair of purely imaginary roots* $\pm iv_0$. *From Eqs. 55.14 and 55.15 we have* $\tau_n = \frac{1}{v_0} \arccos \frac{a_5 v_0 ({v_0}^3 - a_2 v_0) - (a_3 - a_1 {v_0}^2)(a_6 - a_4 {v_0}^2)}{{a_5}^2 {v_0}^2 + (a_6 - a_4 {v_0}^2)^2} + \frac{2j\Pi}{v_0}, \quad j = 1, 2, 3 \ldots$

To show that at $\tau = \tau_0$ there exist a Hopf bifurcation, we need to verify the transversal condition $\frac{d}{d\tau}(Re\rho(\tau))|_{\tau=\tau_0} > 0$. By differentiation (55.12) with respect τ, we get

$$(3\rho^2 + 2a_1\rho + a_2)\frac{d\rho}{d\tau} + [e^{-\rho\tau}(2a_4\rho + a_5) - \tau e^{-\rho\tau}(a_4\rho^2 + a_5\rho + a_6)]\frac{d\rho}{d\tau} = \rho e^{-\rho\tau}(a_4\rho^2 + a_5\rho + a_6). \tag{55.20}$$

From (55.20) we get,

$$\begin{aligned}(\frac{d\rho}{d\tau})^{-1} &= \frac{(3\rho^2 + 2a_1\rho + a_2) + e^{-\rho\tau}(2a_4\rho + a_5) - \tau e^{-\rho\tau}(a_4\rho^2 + a_5\rho + a_6)}{\rho e^{-\rho\tau}(a_4\rho^2 + a_5\rho + a_6)} \\ &= \frac{3\rho^2 + 2a_1\rho + a_2}{\rho e^{-\rho\tau}(a_4\rho^2 + a_5\rho + a_6)} + \frac{2a_4\rho + a_5}{\rho(a_4\rho^2 + a_5\rho + a_6)} - \frac{\tau}{\rho} \\ &= \frac{3\rho^3 + 2a_1\rho^2 + a_2\rho}{-\rho^2(\rho^3 + a_1\rho^2 + a_2\rho + a_3)} + \frac{2a_4\rho^2 + a_5\rho}{\rho^2(a_4\rho^2 + a_5\rho + a_6)} - \frac{\tau}{\rho} \\ &= \frac{2\rho^3 + a_1\rho^2 - a_3}{-\rho^2(\rho^3 + a_1\rho^2 + a_2\rho + a_3)} + \frac{a_4\rho^2 - a_5}{\rho^2(a_4\rho^2 + a_5\rho + a_6)} - \frac{\tau}{\rho}.\end{aligned} \tag{55.21}$$

Thus,

$$\begin{aligned}Sign\{\frac{d(Re\rho)}{d\tau}\}|_{\rho=iv_0} &= Sign\{Re(\frac{d\rho}{d\tau})^{-1}\} \\ &= Sign\{Re[\frac{2\rho^3 + a_1\rho^2 - a_3}{-\rho^2(\rho^3 + a_1\rho^2 + a_2\rho + a_3)}]_{\rho=iv_0} + Re[\frac{a_4\rho^2 - a_6}{\rho^2(a_4\rho^2 + a_5\rho + a_6)}]_{\rho=iv_0} \\ &= Sign\{\frac{2{v_0}^6 + ({a_1}^2 - 2a_2){v_0}^4 - {a_3}^2}{{v_0}^2[(a_1{v_0}^2 - a_3)^2 + ({v_0}^3 - a_2v_0)^2]} + \frac{{a_6}^2 - {a_4}^2{v_0}^4}{{v_0}^2[{a_5}^2{v_0}^2 + (a_6 - a_4{v_0}^2)^2]}\} \\ &= Sign\{\frac{3{v_0}^4 + 2({a_1}^2 - 2a_2 - {a_4}^2){v_0}^2 + ({a_2}^2 - 2a_1a_5 - {a_5}^2 + 2a_4a_6)}{{a_5}^2{v_0}^2 + (a_6 - a_4{v_0}^2)^2}\}.\end{aligned} \tag{55.22}$$

Since v_0 is the largest positive root of the equation (55.16), then we have $\frac{d}{d\tau}(Re\rho(\tau))|_{\tau=\tau_0} > 0$. So, when $\tau > \tau_0$ the real part of $\rho(\tau)$ becomes positive and thus the system $\bar{E}$ becomes unstable. Thus we have the theorem.

Theorem 55.8 *If (i) $R_0 > 1$ (ii) $A_1A_2 - A_3 > 0$ and (iii) $\alpha_3 \le 0, \alpha_2 < 0$ and $\alpha_2 \ge 0$ then the infected steady state is asymptotically stable for all $\tau < \tau_0$ and $\bar{E}$ becomes unstable when $\tau > \tau_0$ where,*

$$\tau_0 = \frac{1}{v_0}\arccos\frac{(a_3 - a_1{v_0}^2)(a_6 - a_4{v_0}^2) - a_5v_0({v_0}^3 - a_2v_0)}{{v_0}^2{a_5}^2 + (a_6 - a_4{v_0}^2)^2}.$$

Hence at $\tau = \tau_0$ Hopf bifurcation occurs, i.e. a family of periodic solution bifurcates for $\bar{E}$ as τ passes through its critical value τ_0. Thus we can comment from the above *Theorem 55.2*, that the delay model reveals Hopf bifurcation at a certain value τ_0 if the parameter are satisfied the condition (i), (ii), (iii). Thus the delay in the disease transmission makes the system stable for the condition stated in *Theorem 55.1* and the system moves towards unstable region for the condition established in *Theorem 55.2*.

55.6.1 Numerical Simulation

In Table 55.1 all parameter values are default values and these values are collected from different Journals [1, 5, 6, 9, 10]. In presence of virus the system is very much inconsistent and thus it is very difficult to choose the parameter values. Many of the parameters have not been measured. They are assumed from different ranges. The new T cells are migrated from thymus at a rate $\lambda = 10\,\text{day}^{-1}\text{mm}^{-3}$ [9]. Since T cells have its natural life span. From this constraint it is assumed that its natural death rate $d_1 = 0.02\,\text{day}^{-1}$ [9, 10]. It has also been observed that the rate of contact between infected and uninfected cell is $0.0025 \sim 0.5$ [11, 12].

Biomedical/virological studies shows that the natural life of infected cell is $2 - 6$ weeks in absence of antigen stimulated replication. We thus assume the natural death rate of infected T cell is $d_2 = 0.24\,\text{day}^{-1}$ in the presence of immune response. In absence of immune impairment effect the stimulation rate of drug induced CTL (i.e. s), rate at which T cell is killed by CTL (i.e. β_2) and natural death rate of drug induced CTL (i.e. d_3) are restricted by the constraint $\frac{s\beta_2}{d_3} \sim 0.01 - 0.05$ [6]. From this restriction we assume the default values of $s, \beta_2, and\, d_3$. We use the default values of the parameter given in Table 55.1. It is also been observed that the another set of parameter for their reported range can be used that give similar behavior.

Figure 55.1 shows that for non delayed system as the cure rate is improved, the uninfected T cell population increases where as the infected T cell and CTL population decreases. Figure 55.2 shows the trajectories for delayed and non delayed system. Here we plot the trajectories for $\tau = 0, \tau = 1$ and $\tau = 5$. This figure shows that as delay is introduced, the system oscillates and as delay is increased the amplitude of oscillation increases. Thus incorporation of delay makes the system unstable.

In Fig. 55.3 we plot a parametric space where β_1 is plotted against s and δ. The relation between these three parameters are derived from the relation $R_0 = \frac{\lambda\beta_1 s}{d_1 d_3(d_2+\delta)}$. Thus the lower region of the figure represents the parametric region for which the steady state E_0 is stable and the upper region represents the region for which E_0 is unstable whereas $\bar{E}$ is asymptotically stable. Figure 55.4 shows the phase plane where β_1 is plotted against s and δ separately. In these two

Table 55.1 Variables and parameters used in the models (55.3) and (55.5)

Parameters	Definition	Default values assigned (day^{-1})
λ	Constant rate of production of $CD4^+T$	10.0 mm^{-3} [9]
d_1	Death rate of uninfected $CD4^+T$ cells	0.01 [9, 10]
β_1	Rate of contact between x and y	0.002 mm^{-3} [12]
d_2	Death rate of virus producing cells	0.24 [9]
β_2	Killing rate of virus producing cells by *CTL*	0.001 mm^{-3} [9]
s	Rate of simulation of *CTL*	0.2 [12]
d_3	Death rate of *CTL*	0.02 [6]
δ	Rate of cure	0.02 [10]

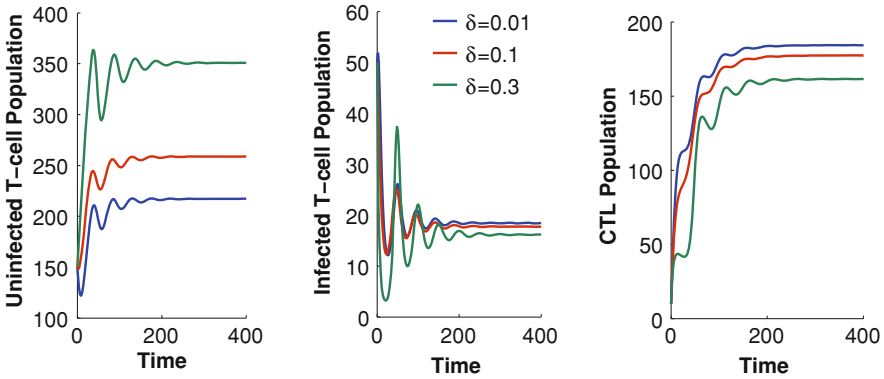


Fig. 55.1 Time series solution of the model variables for different values of δ for non delayed system. Keeping all other parameter as in Table 55.1

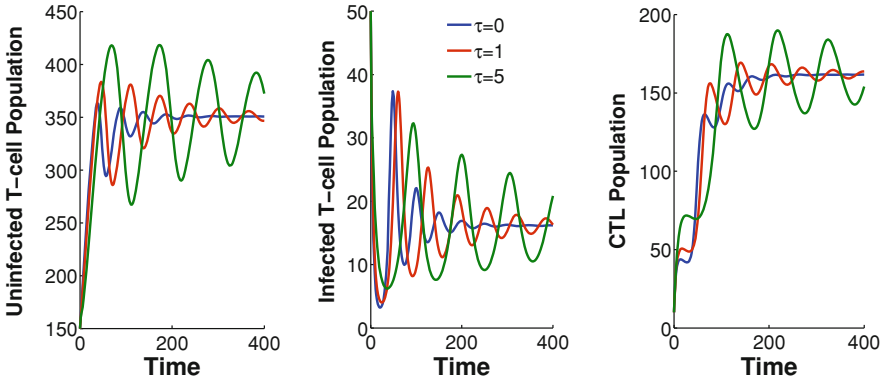


Fig. 55.2 Time series solution of the model variables for delayed and non delayed system. Keeping all other parameter as in Table 55.1

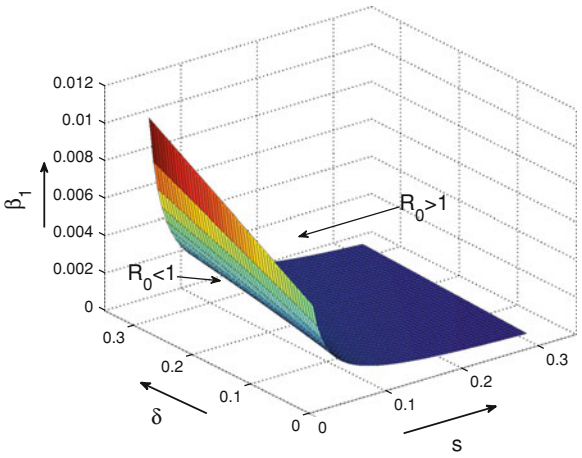


Fig. 55.3 The parametric space for δ-s-β_1. $R_0 < 1$ represents the lower surface region and $R_0 > 1$ represents the upper surface region. Keeping all other parameter as in Table 55.1

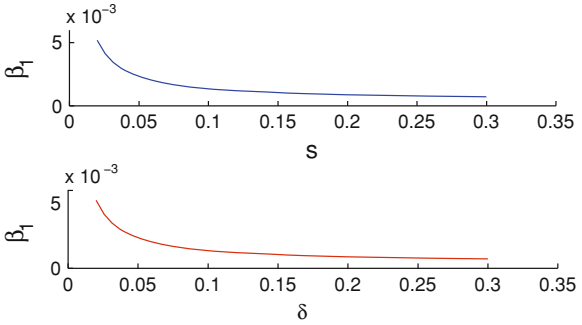


Fig. 55.4 (Top panel) Plot for the diseases transmission rate (β_1) against the rate of stimulation of CTL (s). (Bottom panel) Plot for the diseases transmission rate (β_1) against the cure rate (δ).

cases we see that β_1 decreases as s or δ increases. Thus we can claim that if cure rate or the rate of stimulation is improved then the disease transmission rate can be controlled. Hence for non delayed model if cure rate is improved, diseases transmission would be controlled.

Figs. 55.5 and 55.6 shows that in presence of discrete time delay when δ increases, the uninfected $CD4^+T$ cell population increases whereas the infected T cell and CTL population decreases. Also if λ increases, the rate of uninfected T cell increases rapidly whereas the infected T cell population and CTL population decreases in a slower rate. It is also been observed that in presence of discrete time delay, as λ is increases the system moves to its stable region rapidly.

Figure 55.7 represents the time series solution for different values of δ and τ. In these figure we see that as delay increases from $\tau = 5$ to $\tau = 8$, the amplitude of oscillation of the solution trajectories is increased. Where as if δ is increased from $\delta = 0.02$ to $\delta = 0.8$ the effect of delay is controlled. Thus if the cure rate is

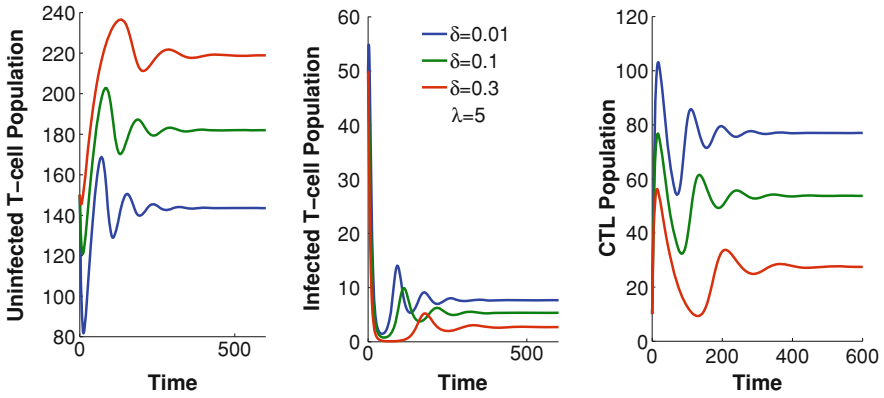


Fig. 55.5 The system behavior for different values of δ when $\lambda = 5$. Keeping all other parameter as in Table 55.1

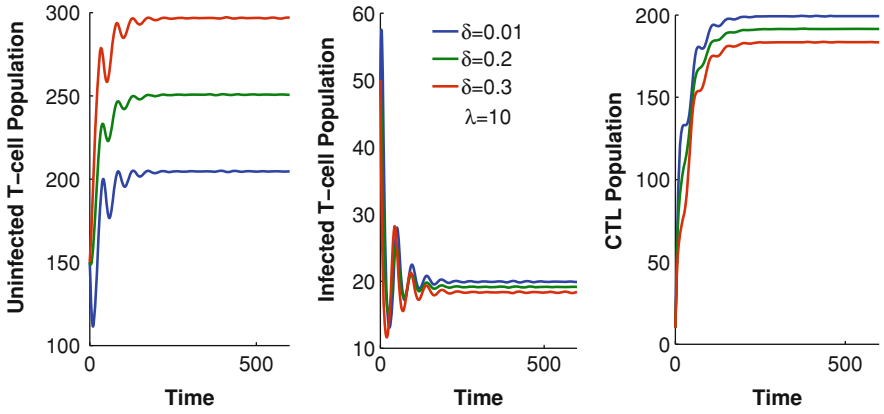


Fig. 55.6 The system behavior for different values of δ when $\lambda = 10$. Keeping all other parameter as in Table 55.1

improved the delay effect can be controlled. Thus the stability of the system is preserved.

55.7 Discussion and Conclusion

Here we presented a mathematical model of HIV infection during long term drug therapy. Again we include a cure rate in the model and delay in the disease transmission term. We have find out the system behavior for the different parameters of the model and also derive some restrictions on parameter such that disease can be controlled. We analyze the existence condition, boundedness and permanence of the system, global stability of the model equations with regard to

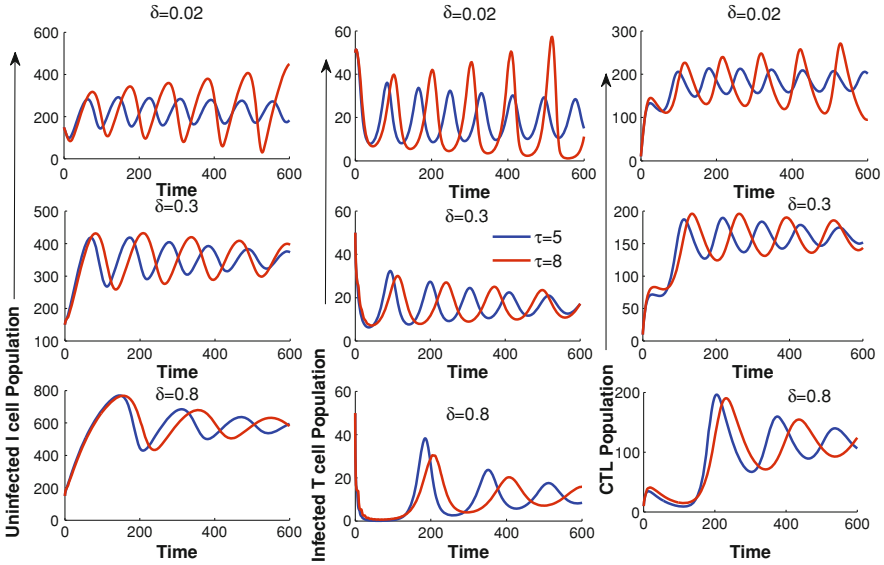


Fig. 55.7 The system behavior for different values of δ when τ. Keeping all other parameter as in Table 55.1

invariance of nonnegativity restriction. We obtain the basic reproduction ratio R_0 which completely determine the dynamics of the system behavior. When $R_0 < 1$, the uninfected steady state is locally stable whereas if $R_0 > 1$ the infected steady state is asymptotically stable for some parametric restriction. We analyze a positive invariant set Γ for which the ODE system is bounded. we also find out that if $R_0 > 1$ the infected equilibrium $\bar{E}$ is globally asymptotically stable. We also studied the delayed system. We incorporate a discrete time delay in the disease transmission term. The condition for which the delay system (i.e. for $\tau > 0$) is stable which is established in Theorem 55.7. Here we see that for small values of time delay the system remains stable under suitable parametric restriction. where as the stability is disrupted through Hopf bifurcation if the delay is increased Theorem 55.8. For this Hopf bifurcation the stability and instability for the delayed system is completely determined. Thus we can find out the existence of region of instability where the population will carry on undergoing habitual vacillation. Numerical simulation confirmed our analytical analysis. Numerically we find that if cure rate or the rate of stimulation is improved then the disease transmission rate can be controlled (Figs. 55.3, 55.4). Hence for non delayed model if cure rate is improved, diseases transmission would be controlled. Our numerical results Fig. 55.7 also shows that increasing value of cure rate could be able to improvise towards stability of the system inspite of delay effect.

Acknowledgments Research is supported by the Government of India, Ministry of Science and Technology, Mathematical Science office, No. SR/S4/MS: 558/08.

References

1. Roy PK, Chatterjee AN, Chattopadhyay B (2010) HIV infection in T-lymphocytes and drug induced CTL response of a time delayed model. In: Lecture notes in engineering and computer science: proceedings of the World Congress on engineering 2010, WCE 2010, 30 June–2 July 2010, London, UK, pp 533–538
2. Carmichael A, Jin X, Sissons P, Borysiewicz L (1993) Quantitative analysis of the human immunodeficiency virus type 1 (HIV-1)-specific cytotoxic T lymphocyte (CTL) response at different stages of HIV-1 infection differential CTL response to HIV-1 and Epstein–Barr virus in late disease. J Exp Med 177:249–256
3. Callaway DS, Perelson AS (2002) HIV-1 infection and low viral loads. Bull Math Biol 64:29–64
4. Coffin JM (1995) HIV population dynamics in vivo implications for genetic variation, pathogenesis, and therapy. Science 267:482–489
5. Kirschner DE, Webb GF (1996) A model of treatment strategy in the chemotherapy of AIDS. Bull Math Biol 58:167–190
6. Bonhoeffer S, Coffin JM, Nowak MA (1997) Human immunodeficiency virus drug therapy and virus load. J Virol 71:3275–3278
7. Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M (1995) Rapid turnover of plasma virons and CD4 lymphocytes in HIV-1 infection. Nature 373:123–126
8. Zhou X, Song X, Shi X (2006) A differential equation model of HIV infection of $CD4^+$T-cells with cure rate. J Math Anal Appl 342:1342–1355
9. Perelson A, Kirschner S, Rob DE, Boer D (1993) Dynamics of HIV infection of $CD4^+$T cells. Math Biosci 114:81–125
10. Culshaw RV, Ruan S, Webb (2003) A mathematical model of cell-to-cell spread of HIV-1 that includes a time delay. Math Biol 46:425–444
11. Nowak MA, May RM (1993) AIDS pathogenis: mathematical models of HIV and SIV infections. AIDS 7:S3–S18
12. Wodarz D, Nowak MA (1999) Specific therapy regimes could lead to long-term immunological control to HIV. Proc Natl Acad Sci USA 96(25):14464–14469